# The American Economic Review

Cuk-H02185-82-P014419

82

## ARTICLES

| | |
|---|---|
| B. S. BERNANKE | Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression |
| G. M. GROSSMAN | Union Wages, Temporary Layoffs, and Seniority |
| W. F. SHUGHART II AND R. D. TOLLISON | Preliminary Evidence on the Use of Inputs by the Federal Reserve System |
| B. SMITH | Limited Information, Credit Rationing, and Optimal Government Lending Policy |
| M. A. AKHTAR | Effects of Interest Rates and Inflation on Aggregate Inventory Investment in the United States |
| L. MAKOWSKI | Competition and Unanimity Revisited |
| S. J. LUNDBERG AND R. STARTZ | Private Discrimination and Social Intervention in Competitive Labor Markets |
| S. WOODWARD | The Liquidity Premium and the Solidity Premium |
| A. E. ROTH AND F. SCHOUMAKER | Expectations and Reputations in Bargaining |
| T. KURAN | Asymmetric Price Rigidity and Inflationary Bias |
| D. A. NICHOLS, D. H. SMALL, AND C. E. WEBSTER, JR. | Why Interest Rates Rise when an Unexpectedly Large Money Stock is Announced |
| R. S. WARREN, JR. | Labor Market Contacts, Unanticipated Wages, and Employment Growth |
| O. J. EVANS | Tax Policy, the Interest Elasticity of Saving, and Capital Accumulation |

SHORTER PAPERS: R. L. Basmann, D. J. Molina, and D. J. Slottje; J. McCallum; W. W. Chang, D. Hamberg, and J. Hirata; G. Loomes and R. Sugden; J. J. Rotemberg; L. S. Seidman; H. C. Wette; J. Gwartney and R. Stroup; E. Katz; A. J. Robson; T. F. Bresnahan; P. Bowbrick; S. Rosen; R. Ram and D. E. Spencer; E. F. Fama; M. H. Ross; R. McKinnon and K.-Y. Tan; S. Edwards; J. A. Hanson; M. L. Weitzman; M. Schwartz and R. J. Reynolds; W. J. Baumol, J. C. Panzar, and R. D. Willig; J. L. Sharma; R. Mendelsohn; V. Tanzi; M. Melvin.

JUNE 1983

# THE AMERICAN ECONOMIC ASSOCIATION

p14419

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1983, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:
$32.00 for regular members with rank of assistant professor or lower, or with annual income of $15,350, or less;
$38.40 for regular members with rank of associate professor, or with annual income of $15,350 to $25,600;
$44.80 for regular members with rank of full professor, or with annual income above $25,600;
$16.00 for junior members (registered students). Certification must be submitted yearly.
Subscriptions (libraries, institutions, or firms) are $100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.
In countries other than the United States, add $9.20 to cover extra postage.

●Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

Founded in 1885

**Officers**

*President*
W. ARTHUR LEWIS
Princeton University

*President-Elect*
CHARLES L. SCHULTZE
The Brookings Institution

*Vice Presidents*
JUANITA M. KREPS
Duke University
EDMUND S. PHELPS
Columbia University

*Secretary*
C. ELTON HINSHAW
Vanderbilt University

*Treasurer*
RENDIGS FELS
Vanderbilt University

*Managing Editor of The American Economic Review*
ROBERT W. CLOWER
University of California-Los Angeles

*Managing Editor of The Journal of Economic Literature*
MOSES ABRAMOVITZ
Stanford University

**Executive Committee**

*Elected Members of the Executive Committee*
ELIZABETH E. BAILEY
Carnegie-Mellon University
ROBERT J. GORDON
Northwestern University
ANN F. FRIEDLAENDER
Massachusetts Institute of Technology
JOSEPH E. STIGLITZ
Princeton University
WILLIAM D. NORDHAUS
Yale University
A. MICHAEL SPENCE
Harvard University

*EX OFFICIO Members*
WILLIAM J. BAUMOL
Princeton University and New York University
GARDNER ACKLEY
The University of Michigan

391
.001

# JOE S. BAIN

## Distinguished Fellow

1982

Joe S. Bain is the undisputed father of modern Industrial Organization Economics. (Edward S. Mason and Edward H. Chamberlin were its two grandparents; but Joe Bain was the father.) His classic text, *Industrial Organization*, published twenty-three years ago, gave the field the rationale and structure that it retains to this day. That approach was already exemplified in his three-volume *Economics of the Pacific Coast Petroleum Industry*, a landmark in the application and empirical testing of the hypotheses of microeconomic theory with respect to the interrelationships between an industry's structure, conduct, and performance.

Bain's theoretical and empirical work on market concentration and the condition of entry, culminating in his "barnbuster," *Barriers to New Competition*, offered the possibility of new, determinate solutions to the oligopoly problem, and added important new insights into the relationship between industry structure, behavior and performance. All of the subsequent burgeoning literature hypothetically elaborating and empirically testing various of these relationships, including, among others, limit pricing and other kinds of strategic behavior, contestability, intraindustry structural and behavioral relationships, shows in both its methods and its substantive theoretical and empirical results the clear imprint of Bain's pioneering work.

# THE AMERICAN ECONOMIC REVIEW

ROBERT W. CLOWER
Managing Editor

JOHN G. RILEY
Associate Editor

WILMA ST. JOHN
Production Editor

THERESA DE MARIA
Assistant Editor

LOIS J. BAGLEY
Editorial Assistant

**Board of Editors**

GEORGE A. AKERLOF
G. O. BIERWAG
THOMAS F. COOLEY
PATRICIA DANZON
RONALD G. EHRENBERG
H. E. FRECH III
JACK HIRSHLEIFER
LAURENCE J. KOTLIKOFF
FREDERIC S. MISHKIN
SHERWIN ROSEN
RICHARD SCHMALENSEE
E. ROY WEINTRAUB

- Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

- No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

- Copyright © American Economic Association 1983. All rights reserved.

**June 1983**

VOLUME 73, NUMBER 3

## Articles

Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression *Ben S. Bernanke* 257

Union Wages, Temporary Layoffs, and Seniority *Gene M. Grossman* 277

Preliminary Evidence on the Use of Inputs by the Federal Reserve System *William F. Shughart II and Robert D. Tollison* 291

Limited Information, Credit Rationing, and Optimal Government Lending Policy *Bruce Smith* 305

Effects of Interest Rates and Inflation on Aggregate Inventory Investment in the United States *M. A. Akhtar* 319

Competition and Unanimity Revisited *Louis Makowski* 329

Private Discrimination and Social Intervention in Competitive Labor Markets *Shelly J. Lundberg and Richard Startz* 340

The Liquidity Premium and the Solidity Premium *Susan Woodward* 348

Expectations and Reputations in Bargaining: An Experimental Study *Alvin E. Roth and Francoise Schoumaker* 362

Asymmetric Price Rigidity and Inflationary Bias *Timur Kuran* 373

Why Interest Rates Rise when an Unexpectedly Large Money Stock is Announced *Donald A. Nichols, David H. Small, and Charles E. Webster, Jr.* 383

Labor Market Contacts, Unanticipated Wages, and Employment Growth *Ronald S. Warren, Jr.* 389

Tax Policy, the Interest Elasticity of Saving, and Capital Accumulation: Numerical Analysis of Theoretical Models *Owen J. Evans* 398

**Shorter Papers**

Budget Constraint Prices as Preference Changing Parameters of Generalized Fechner-Thurstone Direct Utility Functions *R. L. Basmann, D. J. Molina, and D. J. Slottje* 411
Stabilization Policy and Endogenous Wage Stickiness *John McCallum* 414
Liquidity Preference as Behavior Toward Risk is a Demand for Short-Term Securities—Not Money *Winston W. Chang, Daniel Hamberg, and Junichi Hirata* 420
A Rationale for Preference Reversal *Graham Loomes and Robert Sugden* 428
Aggregate Consequences of Fixed Costs of Price Adjustment *Julio J. Rotemberg* 433
Taxes in a Life Cycle Growth Model with Bequests and Inheritances *Laurence S. Seidman* 437
Collateral in Credit Rationing in Markets with Imperfect Information: Note *Hildegard C. Wette* 442
Labor Supply and Tax Rates: A Correction of the Record *James Gwartney and Richard Stroup* 446
Relative Risk Aversion in Comparative Statics *Eliakim Katz* 452
Existence of Consistent Conjectures:
  Comment *Arthur J. Robson* 454
  Reply *Timothy F. Bresnahan* 457
The Economics of Superstars:
  Comment *Peter Bowbrick* 459
  Reply *Sherwin Rosen* 460
Stock Returns, Real Activity, Inflation, and Money:
  Comment *Rati Ram and David E. Spencer* 463
  Reply *Eugene F. Fama* 471
Currency Substitution and Instability in the World Dollar Standard:
  Comment *Myron H. Ross* 473
  Reply *Ronald McKinnon and Kong-Yam Tan* 474
The Short-Run Relation Between Growth and Inflation in Latin America:
  Comment *Sebastian Edwards* 477
  Reply *James A. Hanson* 483
Contestable Markets: An Uprising in the Theory of Industry Structure:
  Comment *Martin L. Weitzman* 486
  Comment *Marius Schwartz and Robert J. Reynolds* 488
  Reply *William J. Baumol, John C. Panzar, and Robert D. Willig* 491
The Choice of Discount Rates for Public Projects:
  Comment *Jandhyala L. Sharma* 497
  Reply *Robert Mendelsohn* 499
Expected Inflation and Interest Rates:
  Comment *Vito Tanzi* 501
  Reply *Michael Melvin* 503

**Auditors' Report** 507

**Notes** 514

# Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression

*By* BEN S. BERNANKE*

During 1930–33, the U.S. financial system experienced conditions that were among the most difficult and chaotic in its history. Waves of bank failures culminated in the shutdown of the banking system (and of a number of other intermediaries and markets) in March 1933. On the other side of the ledger, exceptionally high rates of default and bankruptcy affected every class of borrower except the federal government.

An interesting aspect of the general financial crises—most clearly, of the bank failures—was their coincidence in timing with adverse developments in the macroeconomy.[1] Notably, an apparent attempt at recovery from the 1929–30 recession[2] was stalled at the time of the first banking crisis (November–December 1930); the incipient recovery degenerated into a new slump during the mid-1931 panics; and the economy and the financial system both reached their respective low points at the time of the bank "holiday" of March 1933. Only with the New Deal's rehabilitation of the financial system in 1933–35 did the economy begin its slow emergence from the Great Depression.

A possible explanation of these synchronous movements is that the financial system simply responded, without feedback, to the declines in aggregate output. This is contradicted by the facts that problems of the financial system tended to lead output declines, and that sources of financial panics unconnected with the fall in U.S. output have been documented by many writers. (See Section IV below.)

Among explanations that emphasize the opposite direction of causality, the most prominent is the one due to Friedman and Schwartz. Concentrating on the difficulties of the banks, they pointed out two ways in which these worsened the general economic contraction: first, by reducing the wealth of bank shareholders; second, and much more important, by leading to a rapid fall in the supply of money. There is much support for the monetary view. However, it is not a complete explanation of the link between the financial sector and aggregate output in the 1930's. One problem is that there is no theory of monetary effects on the real economy that can explain *protracted* nonneutrality. Another is that the reductions of the money supply in this period seems quantitatively insufficient to explain the subsequent falls in output. (Again, see Section IV.)

The present paper builds on the Friedman-Schwartz work by considering a third way in which the financial crises (in which we include debtor bankruptcies as well as the failures of banks and other lenders) may have affected output. The basic premise is that, because markets for financial claims are incomplete, intermediation between some classes of borrowers and lenders requires nontrivial market-making and information-gathering services. The disruptions of 1930–33 (as I shall try to show) reduced the effectiveness of the financial sector as a whole in performing these services. As the real costs of intermediation increased, some borrowers (especially households, farmers, and small firms) found credit to be expensive and difficult to obtain. The effects of this credit squeeze on aggregate demand helped convert the severe but not unprecedented downturn of 1929–30 into a protracted depression.

---

*Stanford Graduate School of Business and Hoover Institution. I received useful comments from too many people to list here by name, but I am grateful to each of them. The National Science Foundation provided partial research support.

[1] This is documented more carefully in Sections I.C and IV below.

[2] This paper does not address the causes of the initial 1929–30 downturn. Milton Friedman and Anna Schwartz (1963) have stressed the importance of the Federal Reserve's "anti-speculative" monetary tightening. Others, such as Peter Temin (1976), have pointed out autonomous expenditure effects.

It should be stated at the outset that my theory does not offer a complete explanation of the Great Depression (for example, nothing is said about 1929–30). Nor is it necessarily inconsistent with some existing explanations.[3] However, it does have the virtues that, first, it seems capable (in a way in which existing theories are not) of explaining the unusual length and depth of the depression; and, second, it can do this without assuming markedly irrational behavior by private economic agents. Since the reconciliation of the obvious inefficiency of the depression with the postulate of rational private behavior remains a leading unsolved puzzle of macroeconomics, these two virtues alone provide motivation for serious consideration of this theory.

There do not seem to be any exact antecedents of the present paper in the formal economics literature.[4] The work of Lester Chandler (1970, 1971) provides the best historical discussions of the general financial crisis extant; however, he does not develop very far the link to macroeconomic performance. Beginning with Irving Fisher (1933) and A. G. Hart (1938), there is a literature on the macroeconomic role of inside debt; an interesting recent example is the paper by Frederic Mishkin (1978), which stresses household balance sheets and liquidity. Benjamin Friedman (1981) has written on the relationship of credit and aggregate activity. Hyman Minsky (1977) and Charles Kindleberger (1978) have in several places argued for the inherent instability of the financial system, but in doing so have had to depart from the assumption of rational economic behavior.[5] None of the above authors has emphasized the effects of financial crisis on the real costs of credit intermediation, the focus of the present work.

The paper is organized as follows: Section I presents some background on the 1930–33 financial crisis, its sources, and its correspondence with aggregate output movements. Section II begins the principal argument of the paper. I explain how the runs on banks and the extensive defaults could have reduced the efficiency of the financial sector in performing its intermediary functions. Some evidence of these effects is introduced.

Possible channels by which reduced financial efficiency might have affected output are discussed in Section III. Reduced-form estimation results, reported in Section IV, suggest that augmenting a purely monetary approach by my theory significantly improves the explanation of the financial sector-output connection in the short run. Section V looks at the persistence of these effects.

Some international aspects of the financial sector-aggregate output link are briefly discussed in Section VI and Section VII concludes.

## I. The Financial Collapse: Some Background

The problems faced by the U.S. financial system between October 1930 and March 1933 have been described in detail by earlier authors,[6] but it will be useful to recapitulate some principal facts here. Given this background, attention will be turned to the more central issues of the paper.

The two major components of the financial collapse were the loss of confidence in financial institutions, primarily commercial banks, and the widespread insolvency of debtors. I give short discussions of each of these components and of their joint relation to aggregate fluctuations.

### A. *The Failure of Financial Institutions*

Most financial institutions (even semipublic ones, like the Joint Stock Land Banks) came under pressure in the 1930's. Some, such as the insurance companies and the mutual savings banks, managed to maintain something close to normal operations. Others, like the building-and-loans (which, despite their ability to restrict withdrawals by depositors, failed in significant numbers) were greatly hampered in their attempts to carry on their business.[7] Of most importance, however, were the problems of the commercial banks. The significance of the banking difficulties derived both from their magnitude and from the central role commercial banks played in the financial system.[8]

---

[3] See Karl Brunner (1981) for a useful overview of contemporary theories of the depression. Also, see Robert Lucas and Leonard Rapping's article in Lucas (1981).

[4] This is especially true of the more recent work, which tends to ignore the nonmonetary effects of the financial crisis. Older writers often seemed to take the disruptive impact of the financial breakdown for granted.

[5] I do not deny the possible importance of irrationality in economic life; however, it seems that the best research strategy is to push the rationality postulate as far as it will go.

[6] See especially Chandler (1970, 1971) and Friedman and Schwartz.

The great severity of the banking crises in the Great Depression is well known to students of the period. The percentages of operating banks which failed in each year from 1930 to 1933 inclusive were 5.6, 10.5, 7.8, and 12.9; because of failures and mergers, the number of banks operating at the end of 1933 was only just above half the number that existed in 1929.[9] Banks that survived experienced heavy losses.

The sources of the banking collapse are best understood in the historical context. The first point to be made is that bank failures were hardly a novelty at the time of the depression. The U.S. system, made up as it was primarily of small, independent banks, had always been particularly vulnerable. (Countries with only a few large banks, such as Britain, France, and Canada, never had banking difficulties on the American scale.) The dominance of small banks in the United States was due in large part to a regulatory environment which reflected popular fears of large banks and "trusts"; for example, there were numerous laws restricting branch banking at both the state and national level. Competition between the state and national banking systems for member banks also tended to keep the legal barriers to entry in banking very low.[10] In this sort of environment, a significant number of failures was to be expected and probably was even desirable. Failures due to "natural causes" (such as the agricultural depression of the 1920's upon which many small, rural banks foundered) were common.[11]

Besides the simple lack of economic viability of some marginal banks, however, the U.S. system historically suffered also from a more malign source of bank failures; namely, financial panics. The fact that liabilities of banks were principally in the form of fixed-price, callable debt (i.e., demand deposits), while many assets were highly illiquid, created the possibility of the perverse expectational equilibrium known as a "run" on the banks. In a run, fear that a bank may fail induces depositors to withdraw their money, which in turn forces liquidation of the bank's assets. The need to liquidate hastily, or to dump assets on the market when other banks are also liquidating, may generate losses that actually do cause the bank to fail. Thus the expectation of failure, by the mechanism of the run, tends to become self-confirming.[12]

An interesting question is why banks at this time relied on fixed-price demand deposits, when alternative instruments might have reduced or prevented the problem of runs.[13] An answer is provided by Friedman and Schwartz: They pointed out that, before the establishment of the Federal Reserve in 1913, panics were usually contained by the practice of suspending convertibility of bank deposits into currency. This practice, typically initiated by loose organizations of urban banks called clearinghouses, moderated the dangers of runs by making hasty liquidation unnecessary. In conjunction with the suspension of convertibility practice, the use of demand deposits created relatively little instability.[14]

---

[7] Hart describes the problems of the building-and-loans. An interesting sidelight here is the additional strain on housing lenders caused by the existence of the Postal Savings System; see Maureen O'Hara and David Easley (1979).

[8] According to Raymond Goldsmith (1958), commercial banks held 39.6 percent of the assets of all financial intermediaries, broadly defined, in 1929. See his Table 11.

[9] Cyril Upham and Edwin Lamke (1934, p. 247). Since smaller banks were more likely to fail, the fraction of deposits represented by suspended banks was somewhat less. Eventual recovery by depositors was about 75 percent; see Friedman and Schwartz, p. 438.

[10] Benjamin Klebaner (1974) gives a good brief history of U.S. commercial banking.

[11] Upham and Lamke, p. 247, report that approximately 2–3 percent of all banks in operation failed in each year of the 1920's.

[12] Douglas Diamond and Philip Dybvig (1981) formalize this argument. For an alternative analysis of the phenomenon of runs, see Robert Flood and Peter Garber (1981).

[13] For example, equity-like instruments, such as those used by modern money-market mutual funds, could have been used as the transactions medium. See Kenneth Cone (1982).

However, with the advent of the Federal Reserve (according to Friedman-Schwartz), this roughly stable institutional arrangement was upset. Although the Federal Reserve introduced no specific injunctions against the suspension of convertibility, the clearinghouses apparently felt that the existence of the new institution relieved them of the responsibility of fighting runs. Unfortunately, the Federal Reserve turned out to be unable or unwilling to assume this responsibility.

No serious runs occurred between World War I and 1930; but the many pieces of bad financial news that came in from around the world in 1930–32 were like sparks around tinder. Runs were clearly an important part of the banking problems of this period. Some evidence emerges from contemporary accounts, including descriptions of specific events precipitating runs. Also notable is the fact that bank failures tended to occur in short spasms, rather than in a steady stream (see Table 1, col. 2, for monthly data on the deposits of failing banks). The problem was not arrested until government intervention became important in late 1932 and early 1933.

We see, then, that the banking crises of the early 1930's differed from earlier recorded experience both in magnitude and in the degree of danger posed by the phenomenon of runs. The result of this was that the behavior of almost the entire system was adversely affected, not just that of marginal banks. The bankers' fear of runs, as I shall argue below, had important macroeconomic effects.

### B. *Defaults and Bankruptcies*

The second major aspect of the financial crisis (one that is currently neglected by historians) was the pervasiveness of debtor insolvency. Given that debt contracts were written in nominal terms,[15] the protracted fall in prices and money incomes greatly increased debt burdens. According to Evans Clark (1933), the ratio of debt service to national income went from 9 percent in 1929 to 19.8 percent in 1932–33. The resulting high rates of default caused problems for both borrowers and lenders.

The "debt crisis" touched all sectors. For example, about half of all residential properties were mortgaged at the beginning of the Great Depression; according to the *Financial Survey of Urban Housing* (reported in Hart), as of January 1, 1934,

> The proportion of mortgaged owner-occupied houses with some interest or principal in default was in none of the twenty-two cities [surveyed] less than 21 percent (the figure for Richmond, Virginia); in half it was above 38 percent; in two (Indianapolis and Birmingham, Alabama) between 50 percent and 60 percent; and in one (Cleveland), 62 percent. For rented properties, percentages in default ran slightly higher. [p. 164]

Because of the long spell of low food prices, farmers were in more difficulty than homeowners. At the beginning of 1933, owners of 45 percent of all U.S. farms, holding 52 percent of the value of farm mortgage debt, were delinquent in payments (Hart, p. 138). State and local governments—many of whom tried to provide relief for the unemployed—also had problems paying their debts: As of March 1934, the governments of 37 of the 310 cities with populations over 30,000 and of three states had defaulted on obligations (Hart, p. 225).

In the business sector, the incidence of financial distress was very uneven. Aggregate corporate profits before tax were negative in 1931 and 1932, and after-tax retained earnings were negative in each year from 1930 to 1933 (Chandler, 1971, p. 102). But the subset of corporations holding more than $50 million in assets maintained positive profits throughout this period, leaving the brunt to be borne by smaller companies. Solomon Fabricant (1935) reported that, in 1932 alone, the losses of corporations with assets of $50,000 or less equalled 33 percent of total capitalization; for corporations with assets in the $50,000–$100,000 range, the comparable figure was 14 percent. This led to high rates of failure among small firms.

---

[14] Diamond and Dybvig derive this point formally, with some caveats.

[15] Finding an explanation for the lack of indexed debt during the deflationary 1930's—as in the inflationary 1970's—is a point on which I stumble.

Although the deflation of the 1930's was unusually protracted, there had been a similar episode as recently as 1921–22 which had not led to mass insolvency. The seriousness of the problem in the Great Depression was due not only to the extent of the deflation, but also to the large and broad-based expansion of inside debt in the 1920's. Charles Persons surveyed the credit expansion of the predepression decade in a 1930 article: He reported that outstanding corporate bonds and notes increased from $26.1 billion in 1920 to $47.1 billion in 1928, and that nonfederal public securities grew from $11.8 billion to $33.6 billion over the same period. (This may be compared with a 1929 national income of $86.8 billion.) Perhaps more significantly, during the 1920's, small borrowers, such as households and unincorporated businesses, greatly increased their debts. For example, the value of urban real estate mortgages outstanding increased from $11 billion in 1920 to $27 billion in 1929, while the growth of consumer installment debt reflected the introduction of major consumer durables to the mass market.

Like the banking crises, then, the debt crisis of the 1930's was not qualitatively a new phenomenon; but it represented a break with the past in terms of its severity and pervasiveness.

### C. *Correlation of the Financial Crisis with Macroeconomic Activity*

The close connection of the stages of the financial crisis (especially the bank failures) with changes in real output has been noted by Friedman and Schwartz and by others. An informal review of this connection is facilitated by the monthly data in Table 1. Column 1 is an index of real industrial production. Columns 2 and 3 are the (nominal) liabilities of failing banks and nonbank commercial businesses, respectively.

The industrial production series reveals that a recession began in the United States during 1929. By late 1930, the downturn, although serious, was still comparable in magnitude to the recession of 1920–22; as the decline slowed, it would have been reasonable to expect a brisk recovery, just as in 1922.

With the first banking crisis, however, there came what Friedman and Schwartz called a "change in the character of the contraction" (p. 311). The economy first flattened out, then went into a new tailspin just as the banks began to fail again in June 1931.

A lengthy slide of both the general economy and the financial system followed. The banking situation calmed in early 1932, and nonbank failures peaked shortly thereafter. A new recovery attempt began in August, but failed within a few months.[16] In March 1933, the bottom was reached for both the financial system and the economy as a whole. Measures taken after the banking holiday ended the bank runs and greatly reduced the burden of debt. Simultaneously aggregate output began a recovery that was sustained until 1937.

The leading explanation of the correlation between the conditions of the financial sector and of the general economy is that of Friedman and Schwartz, who stressed the effects of the banking crises on the supply of money. I agree that money was an important factor in 1930–33, but, because of reservations cited in the introduction, I doubt that it completely explains the financial sector-aggregate output connection. This motivates

---

[16] Judging by Table 1, the failure of this recovery seems to be unrelated to financial sector difficulties. However, accounts from the time suggest that the banking crisis of late 1932 and early 1933 (which ended in the banking holiday) was in fact quite severe; see Susan Kennedy (1973). The relatively low reported rate of bank failures at this time may be an artifact of state moratoria, restrictions on withdrawals, and other interventions.

p1449

TABLE 1—SELECTED MACROECONOMIC DATA, JULY 1929–MARCH 1933

| Month | *IP* | *Banks* | *Fails* | $\Delta L/IP$ | $L/DEP$ | *DIF* |
|---|---|---|---|---|---|---|
| 1929*J* | 114 | 60.8 | 32.4 | .163 | .851 | 2.31 |
| *A* | 114 | 6.7 | 33.7 | .007 | .855 | 2.33 |
| *S* | 112 | 9.7 | 34.1 | .079 | .860 | 2.33 |
| *O* | 110 | 12.5 | 31.3 | .177 | .865 | 2.50 |
| *N* | 105 | 22.3 | 52.0 | .121 | .854 | 2.68 |
| *D* | 100 | 15.5 | 62.5 | −.214 | .851 | 2.59 |
| 1930*J* | 100 | 26.5 | 61.2 | −.228 | .837 | 2.49 |
| *F* | 100 | 32.4 | 51.3 | −.102 | .834 | 2.48 |
| *M* | 98 | 23.2 | 56.8 | .076 | .835 | 2.44 |
| *A* | 98 | 31.9 | 49.1 | .058 | .826 | 2.33 |
| *M* | 96 | 19.4 | 55.5 | −.028 | .820 | 2.41 |
| *J* | 93 | 57.9 | 63.1 | .085 | .818 | 2.53 |
| *J* | 89 | 29.8 | 29.8 | −.055 | .802 | 2.52 |
| *A* | 86 | 22.8 | 49.2 | −.027 | .800 | 2.47 |
| *S* | 85 | 21.6 | 46.7 | .008 | .799 | 2.41 |
| *O* | 83 | 19.7 | 56.3 | −.010 | .791 | 2.73 |
| *N* | 81 | 179.9 | 55.3 | −.067 | .777 | 3.06 |
| *D* | 79 | 372.1 | 83.7 | −.144 | .775 | 3.49 |
| 1931*J* | 78 | 75.7 | 94.6 | −.187 | .763 | 3.21 |
| *F* | 79 | 34.2 | 59.6 | −.144 | .747 | 3.08 |
| *M* | 80 | 34.3 | 60.4 | −.043 | .738 | 3.17 |
| *A* | 80 | 41.7 | 50.9 | −.104 | .722 | 3.45 |
| *M* | 80 | 43.2 | 53.4 | −.133 | .706 | 3.99 |
| *J* | 77 | 190.5 | 51.7 | −.120 | .707 | 4.23 |
| *J* | 76 | 40.7 | 61.0 | −.013 | .704 | 3.93 |
| *A* | 73 | 180.0 | 53.0 | −.103 | .706 | 4.29 |
| *S* | 70 | 233.5 | 47.3 | −.050 | .713 | 4.82 |
| *O* | 68 | 471.4 | 70.7 | −.310 | .716 | 5.41 |
| *N* | 67 | 67.9 | 60.7 | −.101 | .726 | 5.30 |
| *D* | 66 | 277.1 | 73.2 | −.120 | .732 | 6.49 |
| 1932*J* | 64 | 218.9 | 96.9 | −.117 | .745 | 4.87 |
| *F* | 63 | 51.7 | 84.9 | −.138 | .757 | 4.76 |
| *M* | 62 | 10.9 | 93.8 | −.183 | .744 | 4.91 |
| *A* | 58 | 31.6 | 101.1 | −.225 | .718 | 6.78 |
| *M* | 56 | 34.4 | 83.8 | −.154 | .696 | 7.87 |
| *J* | 54 | 132.7 | 76.9 | −.170 | .689 | 7.93 |
| *J* | 53 | 48.7 | 87.2 | −.219 | .677 | 7.21 |
| *A* | 54 | 29.5 | 77.0 | −.130 | .662 | 4.77 |
| *S* | 58 | 13.5 | 56.1 | −.091 | .641 | 4.19 |
| *O* | 60 | 20.1 | 52.9 | −.095 | .623 | 4.44 |
| *N* | 59 | 43.3 | 53.6 | −.133 | .602 | 4.79 |
| *D* | 58 | 70.9 | 64.2 | −.039 | .596 | 5.07 |
| 1933*J* | 58 | 133.1 | 79.1 | −.139 | .576 | 4.79 |
| *F* | 57 | 62.2 | 65.6 | −.059 | .583 | 4.09 |
| *M* | 54 | 3276.3[a] | 48.5 | −.767[a] | .607[a] | 4.03 |

*Notes:* *IP* = seasonally adjusted index of industrial production, 1935–39 = 100; *Federal Reserve Bulletin.*

*Banks* = deposits of failing banks, $millions; *Federal Reserve Bulletin.*

*Fails* = liabilities of failing commercial businesses, $millions; *Survey of Current Business.*

$\Delta L/PI$ = ratio of net extensions of commercial bank loans to (monthly) personal income; from *Banking and Monetary Statistics* and *National Income.*

$L/D$ = ratio of loans outstanding to the sum of demand and time deposits, weekly reporting banks; *Banking and Monetary Statistics.*

*DIF* = difference (in percentage points) between yields on Baa corporate bonds and long-term U.S. government bonds; *Banking and Monetary Statistics.*

[a] A national bank holiday was declared in March 1933.

my study of a nonmonetary channel through which an additional impact of the financial crisis may have been felt.

## II. The Effect of the Crisis on the Cost of Credit Intermediation

This paper posits that, in addition to its effects via the money supply, the financial crisis of 1930–33 affected the macroeconomy by reducing the quality of certain financial services, primarily credit intermediation. The basic argument is to be made in two steps. First, it must be shown that the disruption of the financial sector by the banking and debt crises raised the real cost of intermediation between lenders and certain classes of borrowers. Second, the link between higher intermediation costs and the decline in aggregate output must be established. I present here the first step of the argument, leaving the second to be developed in Sections III–V.

In order to discuss the quality of performance of the financial sector, I must first describe the real services that the sector is supposed to provide. The specification of these services depends on the model of the economy one has in mind. We shall clearly not be interested in economies of the sort described by Eugene Fama (1980), in which financial markets are complete and information/transactions costs can be neglected. In such a world, banks and other intermediaries are merely passive holders of portfolios. Banks' choice of portfolios or the scale of the banking system can never make any difference in this case, since depositors can offset any action taken by banks through private portfolio decisions.[17]

As an alternative to the Fama complete-markets world, consider the following stylized description of the economy. Let us suppose that savers have many ways of transferring resources from present to future, such as holding real assets or buying the liabilities of governments or corporations on well-organized exchanges. One of the options savers have is to lend resources to a banking system. The banks also have a menu of different assets to choose from. Assume, however, that banks specialize in making loans to small, idiosyncratic borrowers whose liabilities are too few in number to be publicly traded. (Here is where the complete-markets assumption is dropped.)

The small borrowers to whom the banks lend will be taken, for simplicity, to be of two extreme types, "good" and "bad." Good borrowers desire loans in order to undertake individual-specific investment projects. These projects generate a random return from a distribution whose mean will be assumed always to exceed the social opportunity cost of investment. If this risk is nonsystematic, lending to good borrowers is socially desirable. Bad borrowers try to look like good borrowers, but in fact they have no "project." Bad borrowers are assumed to squander any loan received in profligate consumption, then to default. Loans to bad borrowers are socially undesirable.

In this model, the real service performed by the banking system is the differentiation between good and bad borrowers.[18] For a competitive banking system, I define the *cost of credit intermediation* (*CCI*) as being the cost of channeling funds from the ultimate savers/lenders into the hands of good borrowers. The *CCI* includes screening, monitoring, and accounting costs, as well as the expected losses inflicted by bad borrowers. Banks presumably choose operating procedures that minimize the *CCI*. This is done by developing expertise at evaluating potential borrowers; establishing long-term relationships with customers; and offering loan conditions that encourage potential borrowers to self-select in a favorable way.[19]

Given this simple paradigm, I can describe the effects of the two main components of the financial crisis on the efficiency of the credit allocation process (i.e., on the *CCI*).

---

[17] It should be noted that the phenomena emphasized by Friedman and Schwartz—the effects of the contraction of the banking system on the quantity of the transactions medium and on real output—are also impossible in a complete-markets world.

[18] To concentrate on credit intermediation, I neglect the transactions and other services performed by banks.

[19] See Dwight Jaffee and Thomas Russell (1976) and Joseph Stiglitz and Andrew Weiss (1981) on the way banks induce favorable borrower self-selection.

### A. *Effect of the Banking Crises on the CCI*

The banking problems of 1930–33 disrupted the credit allocation process by creating large, unplanned changes in the channels of credit flow. Fear of runs led to large withdrawals of deposits, precautionary increases in reserve-deposit ratios, and an increased desire by banks for very liquid or rediscountable assets. These factors, plus the actual failures, forced a contraction of the banking system's role in the intermediation of credit.[20] Some of the slack was taken up by the growing importance of alternative channels of credit (see below). However, the rapid switch away from the banks (given the banks' accumulated expertise, information, and customer relationships) no doubt impaired financial efficiency and raised the *CCI*.[21]

It would be useful to have a direct measure of the *CCI*; unfortunately, no really satisfactory empirical representation of this concept is available. Reported commercial loan rates reflect loans that are actually made, not the shadow cost of bank funds to a representative potential borrower; since banks in a period of retrenchment make only the safest and highest-quality loans, measured loan rates may well move inversely to the *CCI*. I obtained a number of interesting results using the yield differential between Baa corporate bonds and U.S. government bonds as a proxy for the *CCI*; however, the use of the Baa rate is not consistent with my story that bank borrowers are those whose liabilities are too few to be publicly traded.

While we cannot observe directly the effects of the banking troubles on the *CCI*, we can see their impact on the extension of bank credit: Table 1 gives some illustrative data. Column 4 gives, as a measure of the flow of bank credit, the monthly change in bank loans outstanding, normalized by monthly personal income.[22] One might have expected the loan-change-to-income ratio to be driven primarily by loan demand and thus by the rate of production. Comparison with the first two columns of Table 2 shows, however, that the banking crises were as important a determinant of this variable as output. For example, except for a brief period of liquidation of speculation loans after the stock market crash, credit outstanding declined very little before October 1930—this despite a 25 percent fall in industrial production that had occurred by that time. With the first banking crisis of November 1930, however, a long period of credit contraction was initiated. The shrinkage of credit shared the rhythm of the banking crises; for example, in October 1931, the worst month for bank failure before the bank holiday, net credit reduction was a record 31 percent of personal income.[23]

The fall in bank loans after November 1930 was not simply a balance sheet reflection of the decline in deposits. Column 5 in Table 1 gives the monthly ratio of outstanding bank loans to the sum of demand and time deposits. This ratio declined sharply as banks switched out of loans and into more liquid investments.

The perception that the banking crises and the associated scrambles for liquidity exerted a deflationary force on bank credit was shared by writers of the time. A 1932 National Industrial Conference Board survey of credit conditions reported that "During 1930, the shrinkage of commercial loans no more than reflected business recession. During 1931 and the first half of 1932 (the period studied), it unquestionably represented pressure by banks on customers for repayment of loans and refusal by banks to grant new loans" (p. 28). Other contemporary sources tended to agree (see, for example, Chandler, 1971, pp. 233–39, for references).

---

[20] For an interesting contemporary account of this process, see the article by Eugene H. Burris in the *American Banker*, October 15, 1931.

[21] Since intermediation resources could have been shifted out of the beleaguered banking sector (given enough time), mine is basically a costs-of-adjustment argument.

[22] In the construction of the bank loans series, data from weekly reporting member banks (which held about 40 percent of all bank loans) were used to interpolate between less frequent aggregate observations. Note that, for our purposes, looking at the change in loans is preferable to considering the stock of real loans outstanding: In a regime of nominally contracted debt and sharp unanticipated deflation, stability of the stock of real debt does *not* signal a comfortable situation for borrowers.

[23] The effect of bank failures on credit outstanding is somewhat exaggerated by the fact that the credit contraction measure includes the loans of suspending banks that were not transferred to other banks; however, I estimate that this accounting convention is responsible for less than one-eighth of the total (measured) credit contraction between October 1930 and February 1933.

Two other observations about the contraction of bank credit can be made. First, the class of borrowers most affected by credit reductions were households, farmers, unincorporated businesses, and small corporations; this group had the highest direct or indirect reliance on bank credit. Second, the contraction of bank credit was twice as large as that of other major countries, even those which experienced comparable output declines (Klebaner, p. 145).

The fall in bank loans outstanding was partly offset by the relative expansion of alternative forms of credit. In the area of consumer finance, retail merchants, service creditors, and nonbank lending agencies improved their position relative to banks and primarily bank-supported installment finance companies (Rolf Nugent, 1939, pp. 114–16). Small firms during this period significantly reduced their traditional reliance on banks in favor of trade credit (Charles Merwin, 1942, pp. 5 and 75). But, as argued above, in a world with transactions costs and the need to discriminate among borrowers, these shifts in the loci of credit intermediation must have at least temporarily reduced the efficiency of the credit allocation process, thereby raising the effective cost of credit to potential borrowers.

### B. *The Effect of Bankruptcies on the CCI*

I turn now to a brief discussion of the impact of the increase in defaults and bankruptcies during this period on the cost of credit intermediation.

The very existence of bankruptcy proceedings, rather than being an obvious or natural phenomenon, raises deep questions of economic theory. Why, for example, do the creditor and defaulted debtor make the payments to third parties (lawyers, administrators) that these proceedings entail, instead of somehow agreeing to divide those payments between themselves? In a complete-markets world, bankruptcy would never be observed; this is because complete state-contingent loan agreements would uniquely define each party's obligations in all possible circumstances, rendering third-party arbitration unnecessary. That we do observe bankruptcies, in our incomplete-markets world, suggests that creditors and debtors have found the combination of simple loan arrangements and *ex post* adjudication by bankruptcy (when necessary) to be cheaper than attempting to write and enforce complete state-contingent contracts.

To be more concrete, let us use the "good borrower-bad borrower" example. In writing a loan contract with a potential borrower, the bank has two polar options. First, it might try to approximate the complete state-contingent contract by making the borrower's actions part of the agreement and by allowing repayment to depend on the outcome of the borrower's project. This contract, if properly written and enforced, would completely eliminate the possibility of either side not being able to meet its obligations; its obvious drawback is the cost of monitoring which it involves. The bank's other option is to write a very simple agreement ("payment of such-amount to be made on such-date"), then to make the loan only if it believes that the borrower is likely to repay. The second approach usually dominates the first, of course, especially for small borrowers.

A device which makes the cost advantage of the simpler approach even greater is the use of collateral. If the borrower has wealth that can be attached by the bank in the event of nonpayment, the bank's risk is low. Moreover, the threat of loss of collateral provides the right incentives for borrowers to use loans only for profitable projects. Thus, the combination of collateral and simple loan contracts helps to create a low effective *CCI*.

A useful way to think of the 1930–33 debt crisis is as the progressive erosion of borrowers' collateral relative to debt burdens.

As the representative borrower became more and more insolvent, banks (and other lenders as well) faced a dilemma. Simple, noncontingent loans faced increasingly higher risks of default; yet a return to the more complex type of contract involved many other costs. Either way, debtor insolvency necessarily raised the *CCI* for banks.

One way for banks to adjust to a higher *CCI* is to increase the rate that they charge borrowers. This may be counterproductive, however, if higher interest charges increase the risk of default. The more usual response is for banks just not to make loans to some people that they might have lent to in better times. This was certainly the pattern in the 1930's. For example, it was reported that the extraordinary rate of default on residential mortgages forced banks and life insurance companies to "practically stop making mortgage loans, except for renewals" (Hart, p. 163). This situation precluded many borrowers, even with good projects, from getting funds, while lenders rushed to compete for existing high-grade assets. As one writer of the time, D. M. Frederiksen, put it:

> We see money accumulating at the centers, with difficulty of finding safe investment for it; interest rates dropping down lower than ever before; money available in great plenty for things that are obviously safe, but not available at all for things that are in fact safe, and which under normal conditions would be entirely safe (and there are a great many such), but which are now viewed with suspicion by lenders. [1931, p. 139]

As this quote suggests, the idea that the low yields on Treasury or blue-chip corporation liabilities during this time signalled a general state of "easy money" is mistaken; money was easy for a few safe borrowers, but difficult for everyone else.

An indicator of the strength of lender preferences for safe, liquid assets (and hence of the difficulty of risky borrowers in obtaining funds) is the yield differential between Baa corporate bonds and Treasury bonds (Table 1, column 6). Because this variable contains no adjustment for the reclassification of firms into higher risk categories, it tends to understate the true difference in yields between representative risky and safe assets. Nevertheless, this indicator showed some impressive shifts, going from 2.5 percent during 1929–30 to nearly 8 percent in mid-1932. (The differential never exceeded 3.5 percent in the sharp 1920–22 recession.) The yield differential reflected changing perceptions of default risk, of course; but note also the close relationship of the differential and the banking crises (a fact first pointed out by Friedman and Schwartz). Bank crises depressed the prices of lower-quality investments as the fear of runs drove banks into assets that could be used as reserves or for rediscounting. This effect of bank portfolio choices on an asset price could not happen in a Fama-type, complete-markets world.

Finally, it is instructive to consider the experience of a country that had a debt crisis without a banking crisis. Canada entered the Great Depression with a large external debt, much of it payable in foreign currencies. The combination of deflation and the devaluation of the Canadian dollar led to many defaults. Internally, debt problems in agriculture and in mortgage markets were as severe as in the United States, while major industries (notably pulp and paper) experienced many bankruptcies (A. E. Safarian, 1959, ch. 7). Although Canadian bankers did not face serious danger of runs, they shifted away from loans to safer assets. This shift toward safety and liquidity, though less pronounced than in the U.S. case, drew criticism from all facets of Canadian society. The *American Banker* of December 6, 1932, reported the following complaint from a nonpopulist Canadian politician:

> The chief criticism of our present system appears to be that in good times credit is expanded to great extremes... but, when the pinch of hard times is first being felt, credit is suddenly and drastically restricted by the banks...At the present time, loans are only being made when the banks have a very wide

> margin of security and every effort is being made to collect outstanding loans. All our banks are reaching out in an endeavor to liquefy their assets....
> [p. 1]

Canadian lenders other than banks also tried to retrench: According to the *Financial Post*, May 14, 1932, "Insurance, trust, and loan companies were increasingly unwilling to lend funds with real estate and rental values falling, a growing number of defaults of interest and principal, the increasing burden of property taxes, and legislation which adversely affected creditors" (quoted in Safarian, p. 130).

More careful study of the Canadian experience in the Great Depression would be useful. However, on first appraisal, that experience does not seem to be inconsistent with the point that even good borrowers may find it more difficult or costly to obtain credit when there is extensive insolvency. The debt crisis should be added to the banking crises as a potential source of disruption of the credit system.

## III. Credit Markets and Macroeconomic Performance

If it is taken as given that the financial crises during the depression did interfere with the normal flows of credit, it still must be shown how this might have had an effect on the course of the aggregate economy.

There are many ways in which problems in credit markets might potentially affect the macroeconomy. Several of these could be grouped under the heading of "effects on aggregate supply." For example, if credit flows are dammed up, potential borrowers in the economy may not be able to secure funds to undertake worthwhile activities or investments; at the same time, savers may have to devote their funds to inferior uses. Other possible problems resulting from poorly functioning credit markets include a reduced feasibility of effective risk sharing and greater difficulties in funding large, indivisible projects. Each of these might limit the economy's productive capacity.

These arguments are reminiscent of some ideas advanced by John Gurley and E. S. Shaw (1955), Ronald McKinnon (1973), and others in an economic development context. The claim of this literature is that immature or repressed financial sectors cause the "fragmentation" of less developed economies, reducing the effective set of production possibilities available to the society.

Did the financial crisis of the 1930's turn the United States into a "temporarily underdeveloped economy" (to use Bob Hall's felicitous phrase)? Although this possibility is intriguing, the answer to the question is probably no. While many businesses did suffer drains of working capital and investment funds, most larger corporations entered the decade with sufficient cash and liquid reserves to finance operations and any desired expansion (see, for example, Friedrich Lutz, 1945). Unless it is believed that the outputs of large and of small businesses are not potentially substitutes, the aggregate supply effect must be regarded as not of great quantitative importance.

The reluctance of even cash-rich corporations to expand production during the depression suggests that consideration of the aggregate demand channel for credit market effects on output may be more fruitful. The aggregate demand argument is in fact easy to make: A higher cost of credit intermediation for some borrowers (for example, households and smaller firms) implies that, for a given *safe* interest rate, these borrowers must face a higher effective cost of credit. (Indeed, they may not be able to borrow at all.) If this higher rate applies to household and small firm borrowing but not to their saving (they may only earn the safe rate on their savings), then the effect of higher borrowing costs is unambiguously to reduce their demands for current-period goods and services. This pure substitution effect (of future for present consumption) is easily derived from the classical two-period model of savings.[24]

---

[24] The classical model may be augmented, if the reader desires, by considerations of liquidity constraints, bankruptcy costs, or risk aversion; see my 1981 paper.

Assume that the behavior of borrowers unaffected by credit market problems is unchanged. Then the paragraph above implies that, for a given safe rate, an increase in the cost of credit intermediation reduces the total quantity of goods and services currently demanded. That is, the aggregate demand curve, drawn as a function of the safe rate, is shifted downward by a financial crisis. In any macroeconomic model one cares to use, this implies lower output and lower safe interest rates. Both of these outcomes characterized 1930–33, of course.

Some evidence on the magnitude of the effect of the financial market problems on aggregate output is now presented.

## IV. Short-Run Macroeconomic Impacts of the Financial Crisis

This section studies the short-run or "impact" effects of the financial crisis. For this purpose, I use only monthly data on the relevant variables. In addition, rather than consider the 1929–33 episode outside of its context, I have widened the sample to include the entire interwar period (January 1919–December 1941).

Section I.C above has already given some evidence of the relationship between the troubles of the financial sector and those of the economy as a whole. However, support for the thesis of this paper requires that nonmonetary effects of the financial crisis on output be distinguished from the monetary effects studied by Friedman and Schwartz. My approach will be to fit output equations using monetary variables, then to show that adding proxies for the financial crisis substantially improves the performance of these equations. Comparison of financial to totally nonfinancial sources of the Great Depression, such as those suggested by Temin, is left to future research.

To isolate the purely monetary influences on the economy, one needs a structural explanation of the money-income relationship. Lucas (1972) has presented a formal model in which monetary shocks affect production decisions by causing confusion about the price level. Influenced by this work, most recent empirical studies of the role of money have related national income to measures of "unanticipated" changes in money or prices.[25]

The most familiar way of constructing a proxy for unanticipated components of a variable is the two-step method of Robert Barro (1978), in which the residuals from a first-stage prediction equation for (say) money are employed as the independent variables in a second-stage regression. I experimented with both the Barro approach and some alternatives.[26] Since my conclusions were unaffected by choice of technique, I report here only the Barro-type results.

In the spirit of the Lucas-Barro analysis, I considered the effects of both "money shocks" and "price shocks" on output. Money shocks ($M-M^e$) were defined as the residuals from a regression of the rate of growth of $M1$ on four lags of the growth rates of industrial production, wholesale prices, and $M1$ itself; price shocks ($P-P^e$) were defined symmetrically.[27] I used ordinary least squares to estimate the effects of money and price shocks on the rate of growth of industrial production, relative to trend.

The basic regression results for the interwar sample period are given as equations (1) and (2) in Table 2. These two equations are of interest, independently of the other results of this paper. The estimated "Lucas supply curve," equation (2), shows an effect of price shocks on output that is statistically and economically significant. As such, it complements the results of Thomas Sargent (1976), who found a similar relationship for the postwar. The relationship of output to money surprises, equation (1), is a bit weaker. The fact that we discover a smaller role for money in the monthly data than does Paul Evans (1981) is primarily the result of our inclusion of lagged values of production on the right-hand side. This inclusion seems justified both on statistical grounds and for

---

[25] A notable exception is Mishkin (1982).

[26] Principal alternatives tried were 1) the use of anticipated as well as unanticipated quantities as explanatory variables; and 2) reestimation of some equations by the more efficient but computationally more complex method of Andrew Abel and Mishkin (1981).

[27] The first-stage regressions were unsurprising and, for the sake of space, are not reported.

TABLE 2—ESTIMATED OUTPUT EQUATIONS

(1)
$$Y_t = \underset{(10.21)}{.623}\,Y_{t-1} - \underset{(-2.37)}{.144}\,Y_{t-2} + \underset{(3.42)}{.407}\,(M-M^e)_t + \underset{(1.16)}{.141}\,(M-M^e)_{t-1}$$
$$+ \underset{(0.42)}{.051}\,(M-M^e)_{t-2} + \underset{(1.19)}{.144}\,(M-M^e)_{t-3}$$

$s.e. = .0272$  $D.W. = 2.02$  Sample: 1/19–12/41

(2)
$$Y_t = \underset{(9.50)}{.582}\,Y_{t-1} - \underset{(-1.76)}{.118}\,Y_{t-2} + \underset{(5.33)}{.533}\,(P-P^e)_t + \underset{(3.33)}{.350}\,(P-P^e)_{t-1}$$
$$+ \underset{(0.34)}{.036}\,(P-P^e)_{t-2} + \underset{(0.66)}{.069}\,(P-P^e)_{t-3}$$

$s.e. = .0260$  $D.W. = 2.01$  Sample: 1/19–12/41

(3)
$$Y_t = \underset{(9.86)}{.613}\,Y_{t-1} - \underset{(-2.63)}{.159}\,Y_{t-2} + \underset{(2.92)}{.332}\,(M-M^e)_t + \underset{(0.99)}{.113}\,(M-M^e)_{t-1} + \underset{(0.96)}{.110}\,(M-M^e)_{t-2}$$
$$+ \underset{(1.38)}{.156}\,(M-M^e)_{t-3} - \underset{(-4.24)}{.869E-04}\,DBANKS_t - \underset{(-1.93)}{.406E-04}\,DBANKS_{t-1}$$
$$- \underset{(-1.95)}{.258E-03}\,DFAILS_t - \underset{(-2.47)}{.325E-03}\,DFAILS_{t-1}$$

$s.e. = .0249$  $D.W. = 1.99$  Sample: 1/21–12/41

(4)
$$Y_t = \underset{(9.76)}{.615}\,Y_{t-1} - \underset{(-2.13)}{.131}\,Y_{t-2} + \underset{(3.99)}{.455}\,(P-P^e)_t + \underset{(1.97)}{.231}\,(P-P^e)_{t-1} - \underset{(-0.03)}{.004}\,(P-P^e)_{t-2}$$
$$+ \underset{(0.22)}{.024}\,(P-P^e)_{t-3} - \underset{(-4.03)}{.799E-04}\,DBANKS_t - \underset{(-1.66)}{.337E-04}\,DBANKS_{t-1}$$
$$- \underset{(-1.52)}{.202E-03}\,DFAILS_t - \underset{(-1.83)}{.242E-03}\,DFAILS_{t-1}$$

$s.e. = .0246$  $D.W. = 1.98$  Sample: 1/21–2/41

*Notes:* $Y_t$ = rate of growth of industrial production (*Federal Reserve Bulletin*), relative to exponential trend.
$(M-M^e)_t$ = rate of growth of $M1$, nominal and seasonally adjusted (Friedman and Schwartz, Table 4-1), less predicted rate of growth.
$(P-P^e)_t$ = rate of growth of wholesale price index (*Federal Reserve Bulletin*), less predicted rate of growth.
$DBANKS_t$ = first difference of deposits of failing banks (deflated by wholesale price index).
$DFAILS_t$ = first difference of liabilities of failing businesses (deflated by wholesale price index).
Data are monthly; $t$-statistics are shown in parentheses.

the economic reason that costs of adjusting production can be presumed to create a serial dependence in output. Like Evans, I was not able to find effects of money (or prices) lagged more than three months.

While these regression results exhibit statistical significance and the expected signs for coefficients, they are disappointing in the following sense: When equations (1) and (2) are used to perform dynamic simulations of the path of output between mid-1930 and the bank holiday of March 1933, they capture no more than half of the total decline of output during the period. This is the basis of the comment in the introduction that the declines in money seem "quantitatively insufficient" to explain what happened to output in 1930–33.

Given the basic regressions (1) and (2), the next step was to examine the effects of including proxies for the nonmonetary financial impact as explanators of output. Based on the earlier analysis of this paper, the most obvious such proxies are the deposits of failing banks and the liabilities of failing businesses.

A preliminary problem with the bank deposits series that needs to be discussed is the value for March 1933, the month of the bank holiday. As can be seen in Table 1, the deposits of banks suspended in March 1933 is seven times that of the next worse month. The question arises if any adjustment should be made to that figure before running the regressions.

We believe that it would be a mistake to eliminate totally the bank holiday episode from the sample. According to contemporary accounts, rather than being an orderly and planned-in-advance policy, the imposition of the holiday was a forced response to the most panicky and chaotic financial conditions of the period. The deposits of suspended banks figure for March, as large as it is, reflects not all closed banks but only those not licensed to reopen by June 30, 1933. Of these banks, most were liquidated or placed in receivership; less than 25 percent had been licensed to reopen as of December 31, 1936.[28] Qualitatively, then, the March 1933 episode resembled the earlier crises; it would be throwing away information not to include in some way the effects of this crisis and of its resolution on the economy.

On the other hand, the mass closing of banks by government action probably created less confusion and fear of future crises than would have a similar number of suspensions occurring without government intervention. As a conservative compromise, I assumed that the "supervised" bank closings of March 1933 had the same effect as an "unsupervised" bank crisis involving 15 percent as much in frozen deposits. This scales down the March 1933 episode to about the size of the events of October 1931. The sensitivity of the results to this assumption is as follows: increasing the amount of importance attributed to the March 1933 crisis raises the magnitude and statistical significance of the measured effects of the financial crises on output. (It is in this sense that the 15 percent figure is conservative.) However, the bank failure coefficients in the regressions retained high significance even when less weight was given to March 1933.

I turn now to the results of adding (real) deposits of failing banks and liabilities of failing businesses to the output equations (see equations (3) and (4) in Table 2). The sample period begins in 1921 because of the unavailability of data on monthly bank failures before then. In both regressions, current and lagged first differences of the added variables enter the explanation of the growth rate of industrial production (relative to trend) with the expected sign and, taken jointly, with a high level of statistical significance. The magnitudes and significance of the coefficients of money and price shocks are not much changed. This provides at least a tentative confirmation that nonmonetary effects of the financial crisis augmented monetary effects in the short-run determination of output.

Some alternative proxies for the nonmonetary component of the financial crisis were also tried. For the sake of space, only a summary of these results is given. 1) To examine the direct effects of the contraction of bank credit on the economy, I began by regressing the rate of growth of bank loans on current and lagged values of suspended bank deposits and of failing business liabilities. (This regression indicated a powerful negative effect of financial crisis on bank loans.) The fitted series from this regression was used as a proxy for the portion of the credit contraction induced by the financial crisis. In the presence of money or price shocks, the effect of a decline in this variable on output was found to be negative for two months, positive for the next two months, then strongly negative for the fifth and sixth months after the decline. For the period from 1921 until the bank holiday, and with monetary variables included, the total effect of credit contraction on output (as measured by the sum of lag coefficients in a polynomial distributed lag) was large (comparable to the monetary effect), negative, and significant at the 95 percent level. For the entire interwar sample, however, the statistical significance of this variable was much reduced. This last result is due to the fact that the recovery of 1933–41 was financed by nonbank sources, with bank loans remaining at a low level.

---

[28] *Federal Reserve Bulletin*, 1937, pp. 866–67.

2) Another proxy for the financial crisis that was tried was the differential between Baa corporate bond yields and the yields on U.S. bonds. As described in Section I.C, this variable responded strongly to both bank crises and the problems of debtors, and as such was a sensitive indicator of financial market conditions. The yield differential variable turned out to enter very strongly as an explanator of current and future output growth, overall and in every subsample. As much of this predictive power was no doubt due to pure financial market anticipations of future output declines, I also put the differential variable through a first-stage regression on the liabilities of bank and business failures. Assuming that these latter variables themselves were not determined by anticipations of future output declines (see below), the use of the fitted series from this regression "purged" the differential variable of its pure anticipatory component. The fitted series entered the output equations less strongly than the raw series, but it retained the right sign and statistical significance at the 95 percent confidence level.

In almost every case, then, the addition of proxies for the general financial crisis improved the purely monetary explanation of short-run (monthly) output movements. This finding was robust to the obvious experiments. For example, with the above-noted exception of the credit variable in 1933–41, coefficients remained roughly stable over subsamples. Another experiment was to include free dummy variables for each quarter from 1931:I to 1932:IV in the above regressions. The purpose of this was to test the suggestion that our results are only a reflection of the fact that both the output and financial crisis variables "moved a lot" during 1930–33. The rather surprising discovery was that the inclusion of the dummies *increased* the magnitude and statistical significance of the coefficients on bank and business failures. Finally, the economic significance of the results was tested by using the various estimated equations to run dynamic simulations of monthly levels of industrial production (relative to trend) for mid-1930 to March 1933. Relative to the pure money-shock and price-shock simulations described above, the equations including financial crisis proxies did well. Equations (3) and (4) reduced the mean squared simulation error over (1) and (2) by about 50 percent. The other (nonreported) equations did better; for example, those using the yield differential variable reduced the *MSE* of simulation from 90 to 95 percent.

These results are promising. However, a caveat must be added: To conclude that the observed correlations support the theory outlined in this paper requires an additional assumption, that failures of banks and commercial firms are not caused by anticipations of (future) changes in output. To the extent that, say, bank runs are caused by the receipt of bad news about next month's industrial production, the fact that bank failures tend to lead production declines does not prove that the bank problems are helping to cause the declines.[29]

While it may not be possible to convince the determined skeptic that bank and business failures are not purely anticipatory phenomena, a good case can be made against that position. For example, while in some cases a bad sales forecast may induce a firm to declare bankruptcy, more often that option is forced by insolvency (a result of past business conditions). For banks, it might well be argued that not only are failures relatively independent of anticipations about output, but that they are not simply the product of current and past output performance either: First, banking crises had never previous to this time been a necessary result of declines in output.[30] Second, Friedman and Schwartz, as well as other writers, have identified specific events that were important sources of bank runs during 1930–33. These include the revelation of scandal at the Bank of the United States (a private bank, which in December 1930 became the largest bank to fail up to that time); the collapse of the Kreditanstalt in Austria and the ensuing financial panics in central Europe; Britain's going off gold; the exposure of huge pyramiding schemes in the United States and Europe; and others, all connected very indirectly (if at all) with the path of industrial production in the United States.

---

[29] Actually, a similar criticism might be made of Barro's work and my own money and price regressions.

[30] Philip Cagan (1965) makes this point; see pp. 216, 227–28. The 1920–22 recession, for example, did not generate any banking problems.

If it is accepted that bank suspensions and business bankruptcies were the product of factors beyond pure anticipations of output decline, then the evidence of this section supports the view that nonmonetary aspects of the financial crisis were at least part of the propagatory mechanism of the Great Depression. If it is further accepted that the financial crisis contained large exogenous components (there is evidence for this in the case of the banking panics), then there are elements of causality in the story as well.

## V. Persistence of the Financial Crisis

The claim was made in the introduction that my theory seems capable, unlike the major alternatives, of explaining the unusual length and depth of the Great Depression. In the previous section, I attempted to deal with the issue of depth; simulations of the estimated regressions suggested that the combined monetary and nonmonetary effects of the financial crisis can explain much of the severity of the decline in output. In this section, the question of the length of the Great Depression is addressed.

As a matter of theory, the duration of the credit effects described in Section II above depends on the amount of time it takes to 1) establish new or revive old channels of credit flow after a major disruption, and 2) rehabilitate insolvent debtors. Since these processes may be difficult and slow, the persistence of nonmonetary effects of financial crisis has a plausible basis. (In contrast, persistence of purely monetary effects relies on the slow diffusion of information or unexplained stickiness of wages and prices.) Of course, plausibility is not enough; some evidence on the speed of financial recovery should be adduced.

After struggling through 1931 and 1932, the financial system hit its low point in March 1933, when the newly elected President Roosevelt's "bank holiday" closed down most financial intermediaries and markets. March 1933 was a watershed month in several ways: It marked not only the beginning of economic and financial recovery but also the introduction of truly extensive government involvement in all aspects of the financial system.[31] It might be argued that the federally directed financial rehabilitation—which took strong measures against the problems of both creditors and debtors—was the only major New Deal program that successfully promoted economic recovery.[32] In any case, the large government intervention is prima facie evidence that by this time the public had lost confidence in the self-correcting powers of the financial structure.

Although the government's actions set the financial system on its way back to health, recovery was neither rapid nor complete. Many banks did not reopen after the holiday, and many that did open did so on a restricted basis or with marginally solvent balance sheets. Deposits did not flow back into the banks in great quantities until 1934, and the government (through the Reconstruction Finance Corporation and other agencies) had to continue to pump large sums into banks and other intermediaries. Most important, however, was a noticeable change in attitude among lenders; they emerged from the 1930–33 episode chastened and conservative. Friedman and Schwartz (pp. 449–62) have documented the shift of banks during this time away from making loans toward holding safe and liquid investments. The growing level of bank liquidity created an illusion (as Friedman and Schwartz pointed out) of easy money;

---

[31] See Chandler (1970), ch. 15, and Friedman and Schwartz, ch. 8.

[32] E. Carey Brown (1956) has argued that New Deal fiscal policy was not very constructive. A paper by Michael Weinstein in Brunner (1981) points out counterproductive aspects of the N.R.A.

however, the combination of lender reluctance and continued debtor insolvency interfered with credit flows for several years after 1933.

Evidence of postholiday credit problems is not hard to find. For example, small businesses, which (as I have noted) suffered disproportionately during the Contraction, had continuing difficulties with credit during recovery. Lewis Kimmel (1939) carried out a survey of credit availability during 1933–38 as a companion to the National Industrial Conference Board's 1932 survey. His conclusions are generally sanguine (this may reflect the fact that the work was commissioned by the American Bankers Association). However, his survey results (p. 65) show that, of responding manufacturing firms normally dependent on banks, refusal or restriction of bank credit was reported by 30.2 percent of very small firms (capitalization less than $50,000); 14.3 percent of small firms ($50,001–$500,000); 10.3 percent of medium firms ($500,001–$1,000,000); and 3.2 percent of the largest companies (capital over $1 million). (The corresponding results from the 1932 NICB survey were 41.3, 22.2, 12.5, and 9.7 percent.)

Two well-known economists, Hardy and Viner, conducted a credit survey in the Seventh Federal Reserve District in 1934–35. Based on "intensive coverage of 2600 individual cases," they found "a genuine unsatisfied demand for credit by solvent borrowers, many of whom could make economically sound use of working capital.... The total amount of this unsatisfied demand for credit is a significant factor, among many others, in retarding business recovery." They added, "So far as small business is concerned, the difficulty in getting bank credit has increased more, as compared with a few years ago, than has the difficulty of getting trade credit." (These passages are quoted in W. L. Stoddard, 1940.)

Finally, another credit survey for the 1933–38 period was done by the Small Business Review Committee for the U.S. Department of Commerce. This study surveyed 6,000 firms with between 21 and 150 employees. From these they chose a special sample of 600 companies "selected because of their high ratings by a standard commercial rating agency." Even within the elite sample, 45 percent of the firms reported difficulty in securing funds for working capital purposes during this period; and 75 percent could not obtain capital or long-term loan requirements through regular markets. (See Stoddard.)

The reader may wish to view the American Bankers Association and Small Business Review Committee surveys as lower and upper bounds, with the Hardy-Viner study in the middle. In any case, the consensus from surveys, as well as the opinion of careful students such as Chandler, is that credit difficulties for small business persisted for at least two years after the bank holiday.[33]

Home mortgage lending was another important area of credit activity. In this sphere, private lenders were even more cautious after 1933 than in business lending. They had a reason for conservatism; while business failures fell quite a bit during the recovery, real estate defaults and foreclosures continued high through 1935.[34] As has been noted, some traditional mortgage lenders nearly left the market: life insurance companies, which made $525 million in mortgage loans in 1929, made $10 million in new loans in 1933 and $16 million in 1934.[35] During this period, mortgage loans that were made by private institutions went only to the very best potential borrowers. Evidence for this is the sharp drop in default rates of loans made in the early 1930's as compared to loans made in earlier years (see Carl Behrens, 1952, p. 11); this decline was too large to be explained by the improvement in business conditions alone.

To the extent that the home mortgage market did function in the years immediately following 1933, it was largely due to the direct involvement of the federal government. Besides establishing some important new institutions (such as the FSLIC and the system of federally chartered savings and loans), the government "readjusted" existing debts, made investments in the shares of thrift institutions, and substituted for recalcitrant private institutions in the provision of direct credit. In 1934, the government-sponsored Home Owners' Loan Corporation made 71 percent of all mortgage loans extended.[36]

---

[33] See Chandler (1970), pp. 150–51.

[34] U.S. Department of Commerce (1975), series N301.

[35] U.S. Department of Commerce (1975), N282.

Similar conditions obtained for farm credit and in other markets, but space does not permit this to be pursued here. Summarizing the reading of all of the evidence by economists and by other students of the period, it seems safe to say that the return of the private financial system to normal conditions after March 1933 was not rapid; and that the financial recovery would have been more difficult without extensive government intervention and assistance. A moderate estimate is that the U.S. financial system operated under handicap for about five years (from the beginning of 1931 to the end of 1935), a period which covers most of the time between the recessions of 1929–30 and 1937–38. This is consistent with the claim that the effects of financial crisis can help explain the persistence of the depression.

## VI. International Aspects

The Great Depression was a worldwide phenomenon; banking crises, though occurring in a number of important countries besides the United States, were not so ubiquitous. A number of large countries had no serious domestic banking problems, yet experienced severe drops in real income in the early 1930's. Can this be made consistent with the important role we have ascribed to the financial crisis in the United States? A complete answer would require another paper; but I offer some observations:

1) The experience of different countries and the mix of depressive forces each faced varied significantly. For example, Britain, suffering from an overvalued pound, had high unemployment throughout the 1920's; after leaving gold in 1931, it was one of the first countries to recover. The biggest problems of food and raw materials exporters were falling prices and the drying up of overseas markets. Thus we need not look to the domestic financial system as an important cause in every case.

2) The countries in which banking crises occurred (the United States, Germany, Austria, Hungary, and others) were among the worst hit by the depression. Moreover, these countries held a large share of world trade and output. The United States alone accounted for almost half of world industrial output in 1925–29, and its imports of basic raw materials and foodstuffs in 1927–28 made up almost 40 percent of the trade in these commodities.[37] The reduction of imports as these economies weakened exerted downward pressure on trading partners.

3) There were interesting parallels between the troubles of the domestic financial system and those of the international system. One of the Federal Reserve's proudest accomplishments had been the establishment, during the 1920's, of an international gold-exchange standard. Unfortunately, like domestic banking, the gold-exchange standard had the instability of a fractional-reserve system. International reserves included not only gold but also foreign currencies, notably the dollar and the pound; for countries other than the United States and the United Kingdom, foreign exchange was 35 percent of total reserves.

In 1931, the expectations that the international financial system would collapse became self-fulfilling. A general attempt to convert currencies into gold drove one currency after another off the gold-exchange standard. Restrictions on the movement of capital or gold were widely imposed. By 1932, only the United States and a small number of other countries remained on gold.

As the fall of the gold standard parallelled domestic bank failures, the domestic insolvency problem had an international analogue as well. Largely due to fixed exchange rates, the deflation of prices was worldwide. Countries with large nominal debts, notably agricultural exporters (the case of Canada has been mentioned), became unable to pay. Foreign bond values in the United States were extremely depressed.

---

[36] U.S. Department of Commerce (1975), N278 and N283.

[37] U.S. Department of Commerce (1947), pp. 29–31.

As in the domestic economy, these problems disrupted the worldwide mechanism of credit. International capital flows were reduced to a trickle. This represented a serious problem for many countries.

To summarize these observations: the fact that the Great Depression hit countries which did not have banking crises does not preclude the possibility that banking and debt problems were important in the United States (or, for that matter, that countries with strong banks had problems with debtor insolvency). Moreover, my analysis of the domestic financial system may be able to shed light on some of the international financial difficulties of the period.

## VII. Conclusion

Did the financial collapse of the early 1930's have real effects on the macroeconomy, other than through monetary channels? The evidence is at least not inconsistent with this proposition. However, a stronger reason for giving this view consideration is the one stated in the introduction: this theory has hope of achieving a reconciliation of the obvious suboptimality of this period with the postulate of reasonably rational, market-constrained agents. The solution to this paradox lies in recognizing that economic institutions, rather than being a "veil," can affect costs of transactions and thus market opportunities and allocations. Institutions which evolve and perform well in normal times may become counterproductive during periods when exogenous shocks or policy mistakes drive the economy off course. The malfunctioning of financial institutions during the early 1930's exemplifies this point.

## REFERENCES

**Abel, Andrew and Mishkin, Frederic,** "An Integrated View of Tests of Rationality, Market Efficiency, and the Short-Run Neutrality of Monetary Policy," Working Paper 726, National Bureau of Economic Research, 1981.

**Barro, Robert,** "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August 1978, *86*, 549–80.

**Behrens, Carl,** *Commercial Bank Activities in Urban Mortgage Financing*, New York: National Bureau of Economic Research, 1952.

**Bernanke, Ben,** "Bankruptcy, Liquidity, and Recession," *American Economic Review Proceedings*, May 1981, *71*, 155–59.

**Brown, E. Carey,** "Fiscal Policy in the 'Thirties: A Reappraisal," *American Economic Review*, December 1956, *46*, 857–79.

**Brunner, Karl,** *The Great Depression Revisited*, Boston: Martinus Nijhoff, 1981.

**Cagan, Philip,** *Determinants and Effects of Changes in the Stock of Money, 1875–1960*, New York: National Bureau of Economic Research, 1965.

**Chandler, Lester,** *America's Greatest Depression*, New York: Harper & Row, 1970.

______, *American Monetary Policy, 1928–1941*, New York: Harper & Row, 1971.

**Clark, Evans,** *The Internal Debts of the United States*, New York: Macmillan Co., 1933.

**Cone, Kenneth,** "Regulation of Depository Financial Intermediaries," unpublished doctoral dissertation, Stanford University, 1982.

**Diamond, Douglas and Dybvig, Philip,** "Bank Runs, Deposit Insurance, and Liquidity," mimeo., University of Chicago, 1981.

**Evans, Paul,** "An Econometric Analysis of the Causes of the Great Depression in the U.S.," mimeo., Stanford University, 1981.

**Fabricant, Solomon,** *Profits, Losses, and Business Assets, 1929–1934*, Bulletin 55, National Bureau of Economic Research, 1935.

**Fama, Eugene,** "Banking in the Theory of Finance," *Journal of Monetary Economics*, January 1980, *6*, 39–57.

**Fisher, Irving,** "The Debt-Deflation Theory of Great Depressions," *Econometrica*, October 1933, *1*, 337–57.

**Flood, Robert and Garber, Peter,** "A Systematic Banking Collapse in a Perfect Foresight World," Working Paper No. 691, National Bureau of Economic Research, 1981.

**Frederiksen, D. M.,** "Two Financial Roads Leading Out of Depression," *Harvard Business Review*, October 1931, *10*, 131–48.

**Friedman, Benjamin,** "Debt and Economic Activity in the United States," Working Paper

No. 704, National Bureau of Economic Research, 1981.

**Friedman, Milton, and Schwartz, Anna J.,** *A Monetary History of the United States, 1867–1960,* Princeton: Princeton University Press, 1963.

**Goldsmith, Raymond,** *Financial Institutions in the American Economy Since 1900,* Princeton: Princeton University Press, 1958.

**Gurley, John G. and Shaw, E. S.,** "Financial Aspects of Economic Development," *American Economic Review,* September 1955, *45,* 515–38.

**Hart, A. G.,** *Debts and Recovery, 1929–1937,* New York: Twentieth Century Fund, 1938.

**Jaffee, Dwight, and Russell, Thomas,** "Imperfect Information and Credit Rationing," *Quarterly Journal of Economics,* November 1976, *90,* 651–66.

**Kennedy, Susan E.,** *The Banking Crisis of 1933,* Lexington: University Press of Kentucky, 1973.

**Kimmel, Lewis H.,** *The Availability of Bank Credit, 1933–1938,* New York: National Industrial Conference Board, 1939.

**Kindleberger, Charles P.,** *Manias, Panics, and Crashes,* New York: Basic Books, 1978.

**Klebaner, Benjamin,** *Commercial Banking in the United States: A History,* Hinsdale: Dryden Press, 1974.

**Lucas, Robert E., Jr.,** "Expectations and the Neutrality of Money," *Journal of Economic Theory,* April 1972, *4,* 103–24.

______, *Studies in Business Cycle Theory,* Cambridge: Massachusetts Institute of Technology Press, 1981.

**Lutz, Friedrich,** *Corporate Cash Balances, 1914–43,* New York: National Bureau of Economic Research, 1945.

**McKinnon, Ronald J.,** *Money and Capital in Economic Development,* Washington: The Brookings Institution, 1973.

**Merwin, Charles L.,** *Financing Small Corporations,* New York: National Bureau of Economic Research, 1942.

**Minsky, Hyman P.,** "A Theory of Systematic Fragility," in E. I. Altman and A. W. Sametz, eds., *Financial Crises,* New York: Wiley-Interscience, 1977.

**Mishkin, Frederic,** "The Household Balance Sheet and the Great Depression," *Journal of Economic History,* December 1978, *38,* 918–37.

______, "Does Anticipated Money Matter? An Econometric Investigation," *Journal of Political Economy,* February 1982, *90,* 22–51.

**Nugent, Rolf,** *Consumer Credit and Economic Stability,* New York: Russell Sage Foundation, 1939.

**O'Hara, Maureen and Easley, David,** "The Postal Savings System in Depression," *Journal of Economic History,* September 1979, *39,* 741–53.

**Persons, Charles E.,** "Credit Expansion, 1920 to 1929 and Its Lessons," *Quarterly Journal of Economics,* November 1930, *45,* 94–130.

**Safarian, A. E.,** *The Canadian Economy in the Great Depression,* Toronto: University of Toronto Press, 1959.

**Sargent, Thomas J.,** "A Classical Macroeconometric Model for the United States," *Journal of Political Economy,* April 1976, *84,* 207–38.

**Stiglitz, Joseph E. and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economic Review,* June 1981, *71,* 393–410.

**Stoddard, W. L.,** "Small Business Wants Capital," *Harvard Business Review,* Spring 1940, *18,* 265–74.

**Temin, Peter,** *Did Monetary Forces Cause the Great Depression?,* New York: W. W. Norton, 1976.

**Upham, Cyril B. and Lamke, Edwin,** *Closed and Distressed Banks: A Study in Public Administration,* Washington: The Brookings Institution, 1934.

**Board of Governors of the Federal Reserve System,** *Banking and Monetary Statistics,* 1943.

**Federal Reserve Bulletin,** various issues.

**National Industrial Conference Board,** *The Availability of Bank Credit,* New York, 1932.

**Survey of Current Business,** various issues.

**U.S. Department of Commerce,** *Historical Statistics of the United States,* Washington: USGPO, 1975.

______, *National Income,* Washington: USGPO, 1954.

______, *The United States in the World Economy,* Washington: USGPO, 1947.

# Union Wages, Temporary Layoffs, and Seniority

*By* Gene M. Grossman*

It is widely believed that the seniority system, in conjunction with the collective bargaining process, contributes significantly to the determination of the unemployment level in unionized sectors. Causal empiricism suggests that real wage demands often remain buoyant in these sectors in the face of seemingly high rates of unemployment among union members. James Medoff convincingly documents the evidence that "adjustment through layoffs is substantially greater in unionized firms than in comparable nonunionized firms" (1979, p. 380). The usual but somewhat informal explanation is that the outcome of union voting is greatly influenced by senior workers who are fairly certain that they will not be laid off when normal shifts in the industry's demand for labor occur.[1] These inframarginal workers are believed to vote for higher wage settlements than they would if their own jobs were not so secure.

Implicit in this argument is an assumption that the size of the union sector is exogenous. In reality, less senior workers are free to exit, and often to enter the unionized sector (albeit at the bottom of the seniority ladder). When less senior workers can express their preferences not only through *voice*, but also by *exit* from the sector, the identity of the more and less senior workers is itself a function of the wage-employment bargain. This suggests that the question of the effect of seniority rules on union wage demands cannot be divorced from that of the endogenous determination of the size of the union sector.

This paper provides a simple, formal model in which the implications of the seniority system can be examined. An important feature of the model is that the union wage rate and the size of the union sector are simultaneously determined. The model adopts an assumption about union behavior, familiar from the works of Wallace Atherton and especially Henry Farber, that follows from a consideration of its internal political process. In particular, the union is assumed to follow a majority voting rule, so that the utility of the median worker (in terms of seniority) is maximized. Since the wage is set for the entire period of the contract, the median worker trades off the extra utility associated with higher wages during periods of employment against the greater probability of being laid off. As mentioned above, the less senior workers can also "vote with their feet." Exit from the union sector is free at the beginning of any contract period. Closed shops are ruled out, so entry occurs whenever there is excess demand for labor in the sector (i.e., relative to the current size of the union). These assumptions imply that, in a steady-state equilibrium, the least senior union member is indifferent between the wage-cum-probability-of-employment package he receives in the union sector and the opportunities available elsewhere in the economy. Together, the hypothesized behavioral relations provide two equations that determine the steady-state union wage and union size, and layoffs in the union sector in every state of nature.

In Section I, this model of a unionized sector is developed, and in Section II, several steady-state comparisons are conducted to investigate how such a sector responds to shifts in the opportunities available outside the unionized sector, shifts in the costs associated with a temporary layoff, and changes in the degree of variability in the value of marginal product of labor in the unionized sector. Section III examines the role of the seniority system itself. A parameter is defined to describe the extent to which a seniority criterion rules in layoff decisions. A general model capable of embodying a full range of assumptions about layoff procedure is specified, and the effects of the seniority system are studied by variation of the key parameter. The final section offers concluding remarks and suggests areas for further research.

---

*Assistant professor, Princeton University. I am grateful to Jean Baldwin Grossman and Carl Shapiro for numerous discussions and thank Orley Ashenfelter, David Card, Ted Cavin, Randy Ellis, Hank Farber, Michael Katz, Andrew Oswald, and Robert Solow for their comments and suggestions.

[1] Medoff and Katharine Abraham provide evidence that senior workers in union firms are more protected from layoffs than similar workers in nonunionized firms. Even the most senior workers are occasionally laid off, however, as, for example, when plants shut down for extended holidays during periods of slack demand.

## I. A Model of a Unionized Sector under Seniority Rule

The production side of a small, unionized sector which operates within a larger economic environment is modelled. The union under consideration is an industrial (as opposed to craft) union. In industries where such unions operate, the firms typically make most employment decisions, and seniority is either the sole or at least a prominent factor in layoff and rehire decisions.[2] Furthermore, it is often the case that any employee who accepts a job in a sector controlled by an industrial union is required to join the union, but that it is illegal to have union membership as a prerequisite for initial employment. These observations will form the basis for some of the assumptions about union-sector behavior made below.

The unionized sector comprises a large number of competitive firms, which can be aggregated to the industry level. The value marginal product of labor (*VMPL*) in the industry during period $t$ is the product of a random element, $\theta_t$, and a nonstochastic element, $g(E_t) > 0$, where $E_t$ is industry employment of the homogeneous workers, and $g'(E) < 0$. The production technology is such that job sharing is not feasible and employment is "all" (eight hours) or "nothing." It is assumed for algebraic convenience that $\theta_t$ is uniformly distributed in the interval $[1-a, 1+a]$, although the basic model would be applicable to more general distributions of the random variable.

I make a standard, but not totally satisfactory assumption, that the union acts as a monopolist, but only with respect to the wage. Thus, the labor demand curve as a function of the state of nature is taken by the union as a constraint. In principle, a union with unlimited monopoly power might exercise it by specifying a take-it-or-leave-it wage and employment contract. Since such packages are rarely observed, they are assumed to be unattainable, with appeal to some limitation on the monopoly power of the union. It would be possible to properly model the bargaining process for a bilateral monopoly, as for example do Ian McDonald and Robert Solow, but the simplification here allows me to focus more sharply on the seniority aspect of wage determination.[3]

Workers in the unionized sector are indexed by $i$, $i \in [0, L_t]$ where $L_t$ is the (endogenous) size of the union at the start of period $t$. Seniority decreases with the index $i$, and it is assumed throughout this section and the next that the sole criterion for layoffs and rehires is seniority. Labor is perfectly mobile *ex ante* between the unionized sector in question and rest of the economy. However, once the state of nature is realized, workers are *ex post* partially immobile, so that laid-off workers cannot accept employment outside the sector during the contract period without bearing some cost.[4] During periods of excess demand, firms can hire nonunion workers, but must pay them the union wage. These workers may join the union prior to the start of the next period, if they choose to do so.[5]

---

[2] Although my model is of course "stylized," a prototypical example of an industrial union might be the United Auto Workers. In contrast to such unions are craft unions, as exemplified by the International Brotherhood of Electrical Workers. In an industry where this type of union operates, hiring decisions are often made by the union, which acts almost as an employment agent. Typically, work is shared, and seniority is not an important consideration.

[3] As McDonald and Solow point out, the monopoly contract as specified here is inefficient, in the sense that all workers and firms could benefit from a fixed employment contract, if firms are risk neutral and workers are risk averse, and if intra-union transfer payments are possible. In particular, the latter assumption will not be made, as discussed below. See also Robert Hall and David Lilien for a discussion of efficient bargains, and an explanation of why unilateral determination of employment by management may arise even in this context. Finally, Martin Weitzman presents a more general analysis of optimal contract type.

[4] The existence of an asymmetry between *ex ante* and *ex post* mobility costs is a common assumption in the

The union objective function has been the subject of much attention of labor economists. For the purpose at hand, it seems essential that explicit recognition be taken of the disparate interests of the various members of the union. If side payments within the union are possible, then the union might maximize the membership's aggregate gain from employment, as in McDonald and Solow. However, if such side payments are impossible, as assumed here, then the union objective function emerges from the political process that governs decision making. Like Farber, I assume this process to be one of majority voting. This yields a median-voter rule; that is, the union acts to maximize the utility of the median voter.[6]

Since labor is homogeneous, no wage gradient would develop in the sector from differences in skills. All union members would nonetheless vote for a wage schedule based on seniority, which, together with the layoff rule, would allow the union to operate as a discriminating monopolist. Such schedules might be costly for the union to negotiate with the firms. The advantage of collective bargaining is that each wage need not be negotiated separately.[7] I henceforth make the strong, but simplifying assumption that all workers receive the same wage.[8]

Given the uniformity of the union wage across workers, it is likely that various state-contingent wage schedules would be favored over a state-independent wage by some but not all union members. Admitting the possibility of state-contingent contracts into the union's internal discussions might make the deliberation process difficult and costly. Furthermore, contingent contracts often entail substantial costs for monitoring and enforcement. For these reasons, and again because such contracts are rarely observed, I summarily rule them out.

Consider, then, the determination of the (state-independent) wage for period $t$, $w_t$, as a function of the size of the union. The probability of employment for a worker with seniority index $i$ is denoted $\pi_i$ and is given by[9]

$$(1) \qquad \pi_i = Pr\{\theta_t g(i) \geqslant w_t\} = Pr\{\theta_t \geqslant w_t/g(i)\} = [1+a-w_t/g(i)]/2a.$$

Let $u(i)$ be the expected utility of a worker with index $i$. An employed worker has in-

---

implicit contracts literature. See, for example, Costas Azariadis or Sanford Grossman and Oliver Hart. If the *ex post* cost of mobility exceeds the difference between the nonunion wage and the value of leisure plus unemployment benefits, then workers who are laid off will remain unemployed for the duration of the contract period. Bengt Holmstrom has shown, in the context of an implicit contracts model encompassing several periods, that workers may choose not to move if they are laid off, even if the current period benefit to doing so is positive. Similarly, if we were to assume in the present context that workers who leave the union sector to accept nonunion employment lose their position in the job queue, or even fail to accumulate seniority during this period, then some workers who are laid off would remain unemployed, rather than leave the union sector, in order to protect their future rents from seniority. Whether or not temporarily laid-off workers remain unemployed until they are rehired or terminate their union membership is immaterial to what follows.

[5] Entry into and exit from the union are discussed further below.

[6] It can be verified below that preferences of union members are single peaked as a function of wage $w$, thus ensuring the existence of a majority-rule equilibrium. See Douglas Blair and David Crawford for discussion of the conditions that guarantee existence and uniqueness of equilibrium in the context of the union voting problem.

[7] Farell Bloch and Mark Kuskin and Richard Freeman show that the union bargaining process results in a standardization of wage rates.

[8] In a model with skill differences, this single wage would represent the benchmark from which adjustments for skill would be made.

[9] More rigorously, $\pi_i$ takes on the value shown in (1) only if it falls between zero and one, and takes on the closer of these extreme values otherwise. Two important restrictions limit the equilibrium value of this function. First, the median worker must have a probability of employment strictly less than one, which implies $1-a-w_t/g(L_t/2)<0$. If this were not so, the median worker would always desire a higher wage. Second, the least senior worker must have a strictly positive probability of employment, i.e., $1+a-w_t/g(L_t)>0$.

come $w_t$, and, suppressing prices and assuming no nonlabor income, indirect utility $v(w_t)$, with $v' > 0$, $v'' < 0$.[10] A laid-off worker achieves utility $\hat{v}$, which is strictly less than the utility $\bar{v}$ associated with *ex ante* employment outside this unionized sector. If temporarily laid-off workers remain unemployed, then $\hat{v}$ combines unemployment benefits and the (fixed) value of leisure. In any event, $\hat{v}$ represents the reservation utility level for all union members. The expected utility of the median worker is

$$(2)\quad u(L_t/2) = \pi_{L_t/2} v(w_t) + \left(1 - \pi_{L_t/2}\right)\hat{v}.$$

The union maximizes this with respect to its choice of $w_t$, which yields a first-order condition

$$(3)\quad [(1+a)g(L_t/2) - w_t]v'(w_t) = v(w_t) - \hat{v},$$

and a second-order condition which is always satisfied. Equation (3) is termed the *voting schedule*, and is shown as curve *VV* in Figure 1. Its slope is

$$\left.\frac{dw}{dL}\right|_{VV} = \frac{(1+a)/2g'(L/2)}{[v(w)-\hat{v}]A/v'(w)+2} < 0,$$

where $A$ is the coefficient of absolute risk aversion, defined to be positive. An increase in the union size results in a less senior median worker who is more concerned with the prospect of being laid off, and therefore votes for a lower wage.

Next, consider the determination of union size. I assume that closed shops (i.e., jobs that have union membership as a prerequisite) are prohibited by law, as in fact they are in the United States. Thus, unions cannot legally prevent firms from hiring nonmembers of the union, if all union members are employed. Entry into the *sector* occurs in any period in which the demand for labor at the going union wage exceeds the union size.

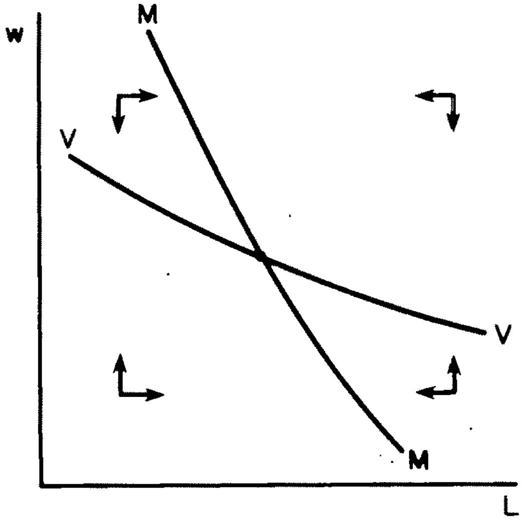


FIGURE 1. EQUILIBRIUM AND STABILITY

The union might nontheless restrict entry into the *union*, so as to achieve a homogeneous voting pool with similar wage objectives. However, unions are also required by law to represent all industry workers in collective bargaining. Membership restriction would entail the loss of membership dues from nonmember employees, who would in effect free ride on the costs of contract negotiation. It is more common, therefore, for unions to negotiate with firms for a membership *requirement* for all labor in the sector, or else to recruit as many members as possible among employees, rather than to restrict membership to certain employees. Thus, it seems plausible to assume that all workers who are offered full-time employment in the union sector in any given period may, if they choose, become union members. Union membership is not an option for workers outside the industry.

Each worker decides whether or not to become or remain a member of the union at the end of each contract period. This decision is based on the expectation that the current period's wage will prevail again in the next period.[11] The benefit to a worker of enlisting or reenlisting in the union is the chance it affords for a high-paying union job. There is a potential cost, however, because *ex post* mobility is costly. If a worker opts to remain in the union sector, and then is laid off after the next period's uncertainty is resolved, then his utility will be less than if he had exited the sector. Let us define $I_t = I(w_{t-1})$ as the seniority index of a worker who, if he were in the sector at the end of period $t-1$, would just be indifferent between remaining in the sector for another period and exiting to the rest of the economy. Then $I_t$ is given implicitly by

---

[10] If $\theta$ is interpreted as the price of output of the unionized sector, then it is assumed that this good is small in the budget set of all consumers.

[11] The assumption of static expectations is for expositional convenience only. The steady-state equilibrium described below is fully consistent with a rational expectations assumption.

$$\pi_{I_t} v(w_{t-1}) + (1-\pi_{I_t})\hat{v} = \bar{v}. \tag{4}$$

Employment in period $t-1$ is given by

$$\theta_{t-1} g(E_{t-1}) = w_{t-1}. \tag{5}$$

Since union exit is free at the end of any period, and entry is open only to those who have worked in the sector, union size in period $t$ is

$$L_t = min\left[I_t, max(L_{t-1}, E_{t-1})\right]. \tag{6}$$

Equations (3)–(6) determine a sequence of wages, employment levels, and union sizes as a function of the realizations of the random variable $\theta$. A steady-state equilibrium of this process, if it exists, must be characterized by

$$L^* = I^*, \tag{7}$$

where asterisks denote steady-state values. To see this, suppose that $L_t < I_t$. Then there are two possibilities. First, if the wage chosen by the union for period $t$ is not "too" high, excess industry demand for labor will result for some values of $\theta_t$. If one of these values is realized, new employees will be hired. Then, either $I_{t+1} > L_t$, and consequently $L_{t+1} > L_t$ (i.e., some of the new workers will elect to join the union), or $I_{t+1} \leqslant L_t$ and therefore $I_{t+1} < I_t$. In either case, the initial situation could not have been one of steady-state equilibrium. Second, the wage may be set by the union at a level sufficiently high to leave a zero probability of excess labor demand. Then the least senior union member will be laid off with probability one, and this worker will certainly not choose to remain in the union when the contract period ends. In such instances, $L_{t+1} < L_t$, and again the initial situation could not have been one of steady-state equilibrium.

Combining (4) and (7) yields the following relationship between the steady-state size of the union and the union wage:

$$\pi_{L^*} v(w^*) + (1-\pi_{L^*})\hat{v} = \bar{v};$$

or, after substituting equation (1),[12]

$$(v(w)-\hat{v})[1+a-(w/g(L))] = 2a(v-\hat{v}). \tag{8}$$

Equation (8) is the *membership schedule*, and is depicted as *MM* in Figure 1. The membership schedule gives combinations of the union wage and union size that leave the least senior union member just indifferent between remaining in the union and seeking employment elsewhere. Note that such an indifference condition is not inconsistent with findings that unions earn premia for holders of union jobs, or that job queues exist in unionized sectors (see John Abowd and Farber). In an *ex ante* sense (i.e., prior to the realization of $\theta$), every union member except the single, least senior worker has higher expected utility than is available outside the union sector.[13] Furthermore, *ex post facto* every worker who does secure a union job earns a wage greater than that prevailing in the nonunion sector. In periods of less than full employment in the unionized sector, the existence of a queue for union jobs (and

---

[12] Henceforth, I shall study only the steady-state equilibria, and will therefore drop the asterisks and time subscripts.

[13] Two referees have suggested that even the marginal worker in the union sometimes enjoys a rent. This is possible, when closed shops are absent, only if union membership is restricted to certain workers among those in the industry. If such restrictions exist, then the question of who is to be admitted to the union is an additional policy issue open to union voting. With two policy variables, a voting equilibrium may fail to exist, except under restrictive assumptions about the sequence of votes.

$$\begin{bmatrix} \frac{v(w)-\hat{v}}{v'(w)}A+2 & -(1+a)g'(L/2)/2 \\ [(1+a)g(L)-w]v'(w)-[v(w)-\hat{v}] & (v(w)-\hat{v})wg'(L)/g(L) \end{bmatrix} \begin{bmatrix} dw \\ dL \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} 0 & 1/v'(w) & g(L/2) \\ 2ag(L) & (1-a)g(L)-w & g(L)[(\bar{v}-\hat{v})+v-v(w)] \end{bmatrix} \begin{bmatrix} d\bar{v} \\ d\hat{v} \\ da \end{bmatrix}.$$

---

perhaps union membership as well) is indeed predicted by the model.

The slope of the *MM* curve is

$$\left.\frac{dw}{dL}\right|_{MM} = \frac{-(v(w)-\hat{v})wg'(L)/g(L)}{[(1+a)g(L)-w]v'(w)-[v(w)-\hat{v}]},$$

which is negative in the neighborhood of the *VV* schedule. An increase in the wage worsens the employment prospects of the least senior worker, lowers his expected utility despite the direct utility increase of the higher wage in employed periods, and, after full adjustment, causes the union sector to shrink.

The equilibrium, steady-state wage, and union size are depicted in Figure 1 at the intersection of the *MM* and *VV* curves. The arrows in the figure indicate the direction of movement of these variables from any initial point. The dynamic system of equations is stable if and only if the *MM* curve is the steeper of the two. I henceforth assume this to be the case.

## II. Steady-State Comparisons

We are now in a position to study how a unionized sector that operates according to a seniority rule responds to various shocks. I concentrate only on comparison of steady-state equilibrium values, and consider three exogenous changes in turn: (*i*) a change in opportunities available outside the unionized sector, $d\bar{v}$; (*ii*) a change in unemployment benefits, the utility of leisure, or mobility costs, represented by $d\hat{v}$; and (*iii*) a mean-preserving change in the variability of the value marginal product of labor in the sector, $da$. The effects of these shocks are derived by totally differentiating equations (3) and (8), and solving the resulting system of equations (9). Note that the Jacobian determinant $\Delta$ is negative by the stability assumption.

### A. *An Increase in $\bar{v}$*

An increase in opportunities outside the unionized sector shifts the *MM* schedule to the left (see Figure 2). At a given wage, and therefore layoff probability, the union must contract in response to an increase in $\bar{v}$. There is no shift in the *VV* curve, since outside opportunities do not directly affect the median worker, who is inframarginal. Thus, the smaller union at point *b* in the figure votes for a higher wage, which causes the union to further contract to point *c*.

At the new equilibrium, the higher wage implies less union employment in any state of nature, and, at the same time, the union labor pool is smaller. It is interesting to consider the net effect on the rate of industry unemployment across states of nature. This variable, which is often mentioned in popular discussions, can be defined as the percentage of workers in an industry who are on temporary layoff, awaiting recall. Of course, since the sector under consideration is small, this is not macroeconomic unemployment in the traditional sense, but stems rather from a labor market distortion at the industry level.[14]

Let $E(\theta)$ be the demand for union labor in state of nature $\theta$, and $z(\theta)=max\{0,[L-E(\theta)]/L\}$ be the industry unemployment

---

[14] In fact, as discussed in fn. 4 above, workers laid off from their jobs in the unionized sector need not remain unemployed while they are awaiting recall.

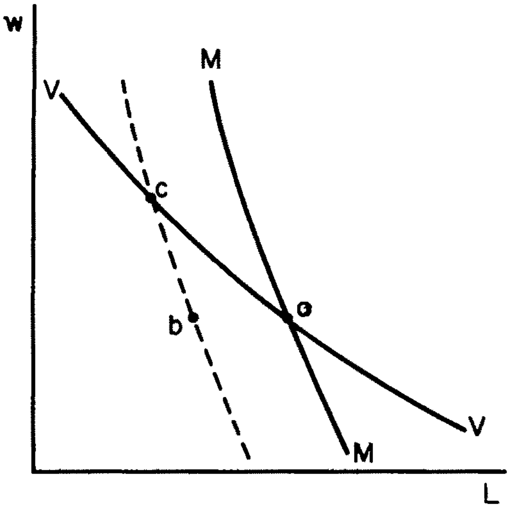


FIGURE 2. INCREASE IN $\bar{v}$

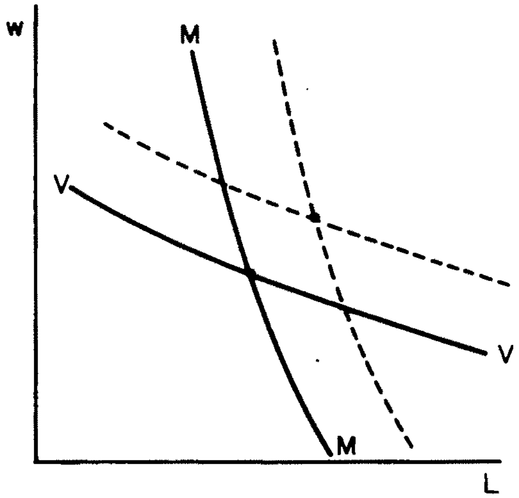


FIGURE 3. INCREASE IN $\hat{v}$

rate. Then for $z(\theta) > 0$,

$$(10)\quad \frac{dz(\theta)}{d\bar{v}} = \frac{2aEg(L)}{L^2\Delta}\left[\frac{(v(w)-\hat{v})}{v'(w)}A + 2 - \frac{(1+a)g(L/2)\varepsilon(L/2)}{w\varepsilon(E)}\right],$$

where $\varepsilon(i) \equiv -ig'(i)/g(i)$, is the elasticity of the *VMPL* schedule, defined to be positive. In general, the effect on the industry unemployment rate is ambiguous, and depends on the curvature of the *VMPL*. Consider the special case in which the elasticity of the *VMPL* is constant (i.e. $\varepsilon(i) = \bar{\varepsilon}$).[15] Substituting (3) into (10), and rearranging, we have

$$sgn\frac{dz(\theta)}{d\bar{v}} = sgn\left\{\frac{v(w)-\hat{v}}{wv'(w)}(1-R)-1\right\},$$

where $r \equiv wA$ is the coefficient of relative risk aversion. *If the VMPL has constant elasticity, then a sufficient condition for the industry unemployment rate to fall in all states of nature in response to an increase in opportunities outside the unionized sector* ($\bar{v}$) *is* $R > 1$. Note that the probability of layoff must rise for all workers who remain in the union, implying that utility falls for all remaining workers with seniority less than the median.

### B. *An Increase in* $\hat{v}$

An increase in unemployment benefits or a decrease in mobility costs shifts both the *MM* and *VV* schedules to the right, as shown in Figure 3.[16] For a given wage, an increased utility associated with layoffs implies a larger union. For a given union size, an increased $\hat{v}$ induces the median voter to support a higher wage. With these shifts, both the wage and the union size can move in either direction, although both variables cannot be observed to decrease together.

The solution for the change in union size can be derived from (9), and rearranged to yield

---

[15] A constant-elasticity *VMPL* schedule corresponds to an underlying production function that is Cobb-Douglas.

[16] It is assumed here that the change in unemployment benefits or mobility costs does not affect the utility level achievable outside the unionized sector. For a complete treatment, we would need to model the remainder of the economy, in order to calculate $d\bar{v}/d\hat{v}$.

$$(11)\quad \frac{dL}{d\hat{v}}=\frac{1}{\Delta}\{[(1-3a)g(L)-w] + [(1+a)g(L/2)-w] + A[(1+a)g(L/2)-w] \times[(1-a)g(L)-w]\}.$$

The last term in the brackets is negative, as is $\Delta$. If $a<1/3$, then the sum of the first two terms in the brackets is negative as well. Thus, for relatively little variability in the *VMPL* ($a$ small), $dL/d\hat{v}>0$. If $a$ is large, the union will nonetheless expand in size when unemployment benefits increase, if the coefficient of absolute risk aversion is sufficiently large. A sufficient condition for $dL/dv>0$ when $a>1/3$ is $A>1/[w-(1-a)g(L)]$. A fall in union size is possible only for a large shift in the *VV* schedule, and a small shift in the *MM* schedule. The former occurs if absolute risk aversion is small or utility associated with temporary layoff is great. The latter occurs if the probability of employment for the least senior worker is high or if the *VMPL* evaluated at $L$ is very employment elastic.

The change in the wage caused by a change in the utility of a temporarily laid-off worker is

$$(12)\quad dw/d\hat{v}=[-2ag(L/2)/L\Delta] \times\left[(1+a)g(L)\times\varepsilon(L/2)(\pi_L-1) + w\varepsilon(L)\pi_{L/2}\right].$$

This is likely to be positive when the probabilities of employment for both the median and least senior worker are high, which in turn occurs whenever unemployment benefits are initially low and mobility costs high, and when opportunities outside the unionized sector are favorable.

The expression for the change in the industry unemployment rate for a general $g(L)$ function is complicated, and of ambiguous sign. Again I consider the special case in which the elasticity of the *VMPL* schedule is constant. For this case, the following expressions can be derived:

$$(13)\quad \frac{dz(\theta)}{d\hat{v}}=\frac{E}{L^2\Delta}\left\{-2ag(L)+(R-1) \times[(1-a)g(L)-w]\left[\frac{v(w)-\hat{v}}{wv'(w)}\right]\right\}.$$

A sufficient condition for (13) to be positive is $R>1$. *If the VMPL schedule has constant elasticity and the coefficient of relative risk aversion exceeds unity, then a rise in unemployment benefits (or a fall in mobility costs) increases the unemployment rate in the unionized sector in all states of nature with positive unemployment.*

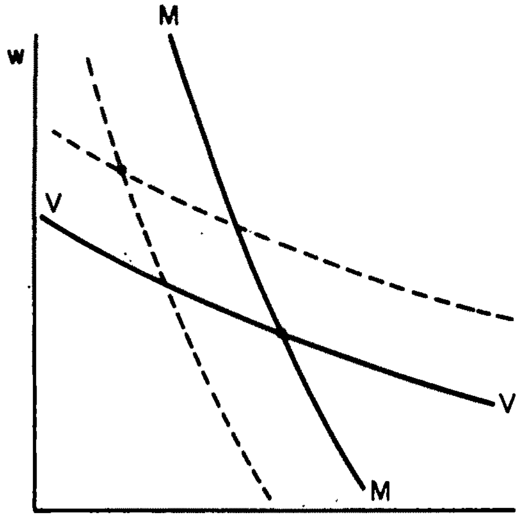
### C. *An Increase in $\underline{a}$*

Finally, I consider a mean-preserving spread in the random, multiplicative shift factor in the *VMPL*, as represented by an increase in $a$. This can be regarded loosely as (the negative of) the effect of a government "leaning against the wind" policy with respect to industry-specific fluctuations (for example, import competition).

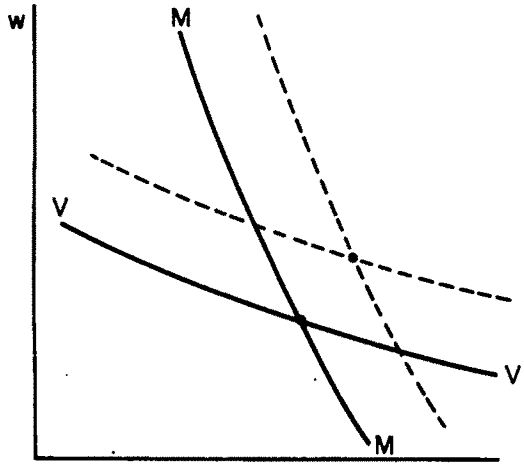
An increase in $a$ shifts the *VV* schedule to the right, as depicted in both a and b of Figure 4. For fixed union size, the change in the most-preferred wage of the median worker that is caused by a change in the range of $\theta$ is given by

$$\left.\frac{dw}{da}\right|_{VV,dL=0}=\frac{g(L/2)}{(v(w)-\hat{v})A/v'(w)+2}>0.$$

The direction of shift of the *MM* curve depends upon how the probability of employment of the initial least senior worker is affected by a spread in $\theta$. If the least senior worker is initially employed with probability greater than one-half, then the probability of employment falls as $a$ increases, and the *MM* curve shifts to the left. This case, which is certainly the most plausible on empirical grounds, is illustrated in Figure 4a. If instead $\pi_L<1/2$ initially, then the *MM* curve shifts to the right, as in Figure 4b. The magnitude



a: $\pi_L > 1/2$



b: $\pi_L < 1/2$

FIGURE 4. INCREASE IN $a$

of the shift is

$$\left.\frac{dL}{da}\right|_{MM,dw=0}=\frac{[g(L)-w]L}{aw\varepsilon(L)}.$$

Note that $g(L)>w \Leftrightarrow \pi_L>1/2$, so that the direction of shift of the $MM$ curve is indeed governed by the size of the initial probability of employment of the least senior worker.

Consider the more likely case in which $\pi_L>1/2$ (Figure 4a). In this case, unions that operate under a seniority system will *ceteris paribus* be smaller in industries subject to wider fluctuations, and union wages will be higher in these industries.[17] This suggests that a government policy aimed at stabilizing a unionized sector may actually effect an increase in the average layoff rate in the industry, as a result of the induced expansion of the industry's labor pool. In the special case of a constant elasticity $VMPL$ curve, it can be shown that

$$\begin{aligned}&sgn\{dz(\theta)/da\}\\&\quad=sgn\{(v(w)-\hat{v})(1-R)/wv'(w)-1\}.\end{aligned}$$

A necessary and sufficient condition for the unemployment rate to increase in every state of nature for which unemployment is positive (given $\pi_L>1/2$ and $\varepsilon(i)=\bar{\varepsilon}$), in response to a government stabilization program, is $R>1-wv'(w)/(v(w)-\hat{v})$. When workers are sufficiently risk averse, a dampening of industry fluctuations induces many more workers to enter the sector, and causes the layoff rate to rise, despite the lower wage and consequent higher labor demand.

To summarize, the simple model of union behavior and industry equilibrium yields relatively simple and plausible answers to questions concerning the effects on steady-state equilibrium of exogenous shocks to the unionized sector. In all cases the coefficient of relative risk aversion plays an important role, because it summarizes the extent to which junior workers are willing to gamble on the possibility of landing a high-paying union job. If workers are relatively risk averse (for example, $R>1$), and the $VMPL$ has constant elasticity, then the industry unemployment rate will increase when opportunities outside the union sector worsen, when unemployment benefits improve, and when fluctuations in the sector's value marginal product of labor are dampened.

---

[17] For $\pi_L<1/2$, a sufficient condition for an increase in $a$ to nonetheless result in a higher union wage is $\varepsilon(L)>\varepsilon(L/2)$, i.e., that the elasticity of the $VMPL$ schedule be a nondecreasing function of employment.

## III. The Effect of the Seniority Rule for Layoffs

Does the seniority criterion for layoffs and rehires cause the union wage and the industry unemployment rate to be higher than would result if workers were unsure of their place in the job queue? In order to address this question, we must extend the model of the previous section to allow for uncertainty on the part of each worker regarding his location in the layoff ordering. The purpose of the present section is to develop one such extension which introduces a single parameter that summarizes the degree of worker uncertainty. Once this is done, we can study the effects of the seniority system by varying this important parameter.

The basic structure of the above model is adopted. In particular, the representation of the stochastic *VMPL* schedule, the nature of the relationship between the union and the firms in the industry, and the modus operandi of the union are left unchanged. However, workers are no longer assumed to be perfectly certain about the order in which layoffs will take place. Let $i \in [\delta, L-\delta]$ once again be an index for members of the union, ranked according to decreasing seniority, and let there be a uniform density of $L/(L-2\delta)$ workers of each type. The total membership of the union is again $L$. A union member with index $i$ has a subjective prior for his location in the job queue that is uniformly distributed on $[i-\delta, i+\delta]$. Workers with higher indices expect on average to be laid off sooner. Note that for $\delta = 0$, this specification of workers' expectations and the distribution of types of workers is the same as that in Section I. When $\delta$ takes on its maximum allowable value of $L/2$, all workers are identical, and each expects that layoff decisions will be made randomly, and that seniority considerations are completely inoperative.[18]

The subjective probability of employment for a worker with index $i$ is

$$(14)\quad \pi_i = \int_{i-\delta}^{i+\delta} [(1+a) - w/g(j)] \frac{1}{4a\delta} dj,$$

where $j$ is the potential ranking of worker $i$ in the job queue. We can now proceed as before. The median worker with index $i = L/2$ has expected utility $u(L/2) = \pi_{L/2} v(w) + (1-\pi_{L/2})\hat{v}$. The union voting schedule is derived by maximizing $u(L/2)$ with respect to the choice of the wage rate. The equation for the *VV* curve, analogous to equation (3) above, is

$$(15)\quad \int_{L/2-\delta}^{L/2+\delta} [1+a-w/g(j)] \frac{1}{\delta} dj = \frac{v(w)-v}{v'(w)} \int_{L/2-\delta}^{L/2+\delta} \frac{1}{\delta g(j)} dj.$$

The *MM* curve, which applies in a steady-state equilibrium, equates the utility of the least senior worker (whose index is $i = L-\delta$) with the utility level attainable outside the unionized sector. Algebraically, this is represented as $\pi_{L-\delta} v(w) + (1-\pi_{L-\delta})\hat{v} = \bar{v}$, or

$$(16)\quad \left[\int_{L-2\delta}^{L} \{1+a-w/g(j)\} \frac{1}{4a\delta} dj\right] \times [v(w)-\hat{v}] = \bar{v}-\hat{v}.$$

It is straightforward to show that both of these curves must be downward sloping (in the neighborhood of steady-state equilibrium), and that the *MM* curve must be the steeper of the two for stability. Once again, this will be assumed to hold throughout.

An increase in $\delta$ represents an increase in the extent of uncertainty about position in the job queue. If the seniority ranking were the only criterion for layoffs and rehires, no such uncertainty could exist. Therefore, an increase in $\delta$ implies that other considerations besides seniority are becoming more important in these decisions. I use $\delta$ as a measure of the strictness of reliance on the seniority system, and investigate the role of seniority by differentiating (15) and (16) totally with respect to $\delta$.

---

[18] This is the specification adopted by McDonald and Solow, who, of course, are not concerned with the role of seniority.

An increase in $\delta$ causes the *MM* curve to shift to the right. This is because a reduction in reliance on the seniority system must necessarily raise the probability of employment of the least senior worker. This result is more general than the specific representation of the workers' expectations used here, since the least senior worker's place on the job queue obviously can be no worse than last. At a given wage, movement away from the seniority system results in an expansion in steady-state union size given by

$$\left.\frac{dL}{d\delta}\right|_{MM,\,dw=0} = \frac{1/g(L-2\delta)-1/g(L)}{\dfrac{2}{g(L-2\delta)}-\displaystyle\int_{L-2\delta}^{L}\frac{1}{\delta g(j)}\,dj}>0.$$

The direction of shift of the *VV* schedule depends critically on how the probability of employment of the median worker is affected by relaxation of the seniority rule. The probability of employment of the median worker decreases with $\delta$ if and only if the function $h(E)\equiv 1/g(E)$ is convex (i.e., $h''>0$). This is better understood with the aid of Figure 5. Here I plot the probability of employment of the median (in terms of seniority) worker, $\pi_{L/2}=[1+a-wh(j)]/2a$, as a function of $j$, the possible position of the median worker in the job queue. If this function is convex as drawn (as will be the case if $h(E)$ is concave), then a gamble drawn from a uniform distribution between $j=L/2-\delta$ and $j=L/2+\delta$ must raise the mean subjective probability of employment for the median worker, relative to a certain employment probability at $j=L/2$.

When is $h(E)$ likely to be convex? Consider the *VMPL* schedule that is derived from a constant returns to scale, constant elasticity of substitution (*CES*) production function for an industry with a fixed stock of capital.[19] Let $\sigma$ be the elasticity of substitution and $\theta$ the elasticity of output with respect to labor. Then a necessary and sufficient condition for $h(E)$ convex is $\sigma<1/(1+\theta)$.

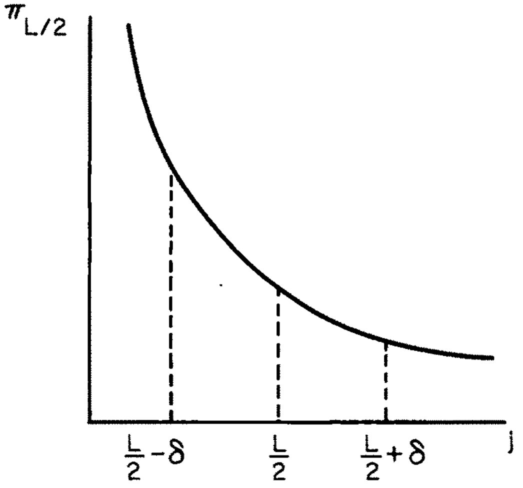


FIGURE 5. EMPLOYMENT PROSPECTS FOR THE MEDIAN WORKER

We can now express the change in the outcome of the union-wage vote (for a given union size), in response to a relaxation of the seniority rule for layoffs as

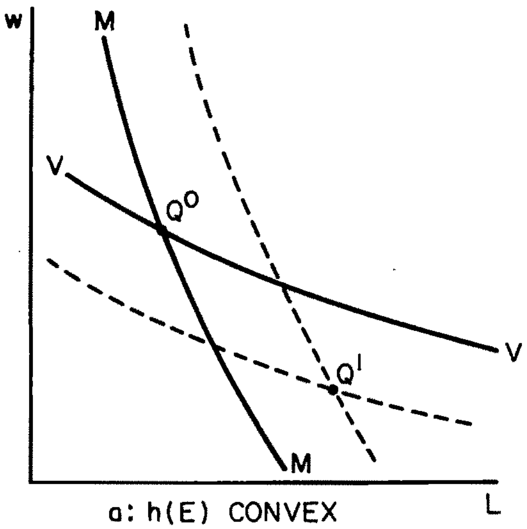
$$\left.\frac{dw}{d\delta}\right|_{VV,\,dL=0}=\Bigg[[wv'(w)+v(w)-v]\times\left[1\Big/g\left(\frac{L}{2}+\delta\right)+1\Big/g\left(\frac{L}{2}-\delta\right)-\int_{L/2-\delta}^{L/2+\delta}\frac{1}{g(j)}\,dj\right]\Bigg]\div\left[\int_{L/2-\delta}^{L/2+\delta}\{v''(w)[1+a-w/g(j)]-2v'(w)/g(j)\}\,dj\right].$$

The denominator is negative, and the first bracketed term in the numerator is positive, so the wage increases (decreases) if the second bracketed term is negative (positive), or if $h(E)$ is concave (convex).

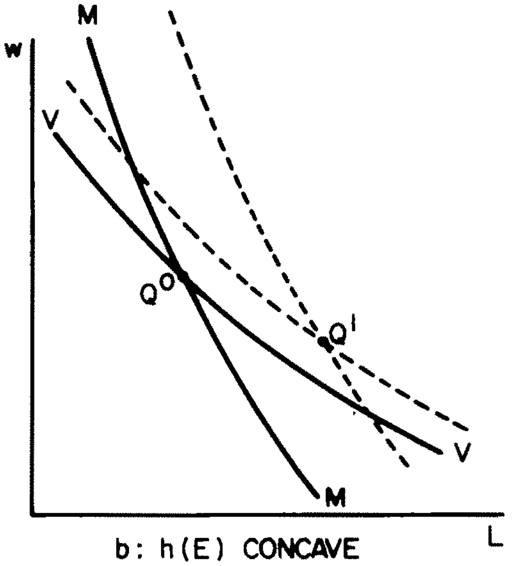
Consider first the case in which $h(E)$ is convex. The result of a parametric increase

---

[19] That is, $g(L)=\partial F(L;K)/\partial L$, where $F(L,K)=[\alpha L^{(\sigma-1)/\sigma}+(1-\alpha)K^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)}$.



a: h(E) CONVEX



b: h(E) CONCAVE

FIGURE 6. INCREASE IN $\delta$

in $\delta$ in this case is depicted in Figure 6a. The *VV* curve has been shifted downward, because at the initial equilibrium $Q^0$, the median worker finds that the mean-preserving spread in his probable position in the job queue has reduced his subjective probability of employment. He therefore readjusts his most-favored wage downward, to once again achieve an optimal point on his wage vs. probability-of-employment tradeoff. At the same time, new members join the union, initially because the move away from the seniority rule improves their prospects, and then in response to the new, lower, union wage. The ultimate steady-state equilibrium is at $Q^1$. Here the seniority system unambiguously has two of the effects usually attributed to it. A reduction in the use of the seniority criterion for layoffs results in both a lower wage rate and a larger union. The direction of change of the industry unemployment rate is less clear, since both the labor pool and labor demand grow. The average unemployment rate is most likely to fall when the coefficient of relative risk aversion is small, since the decline in the wage is largest in such circumstances.

If $h(E)$ is concave, as would be the case if the elasticity of substitution between capital and labor for a CES production function were large, then increased uncertainty about how seniority translates into position in the job queue raises the subjective probability of employment for the median worker. It is perhaps surprising that in such cases the seniority system, in conjunction with the majority voting process, *mitigates* the wage demands of the union. The increase in uncertainty about layoff order now shifts the *VV* curve upward, as in Figure 6b. The wage rate may nonetheless fall, as in the case illustrated, but the reason is not the one that is commonly offered. Here the wage falls not as a result of the voting behavior of a given median worker, but because the enhanced opportunity for less senior workers causes the union to expand, and alters the *identity* of the median worker.

For larger shifts in the *VV* schedule than the one shown, a movement away from a seniority rule could *increase* the wage rate. Shifts in the *VV* schedule are large when the coefficient of relative risk aversion is small, and when the elasticity of substitution between labor and the fixed factors in the industry is large. It is straightforward but tedious to show that for a constant elasticity *VMPL* schedule, as would derive from a Cobb-Douglas production function, an increase in $\delta$ unambiguously results in a higher wage. Such a wage increase could possibly occur concomitant with an expansion in the size of the unionized sector, in which case the industry unemployment level would necessarily rise in every state of nature. In general, not much more can be said about the case of $h(E)$ concave.

Let us recapitulate the arguments of this section. The received wisdom concerning the combined role of the seniority system and majority voting within unions in the determination of union wages must implicitly rely on an assumption that movements away from a seniority criterion for layoffs decrease the probability of employment of the median worker. I have shown that this is in fact true when the reciprocal of the value marginal product of labor schedule is a convex function of employment. A simple necessary and sufficient condition was derived for the example of a *CES* production function. When the condition is satisfied, a movement toward a stricter seniority system, in the sense that seniority becomes a better predictor of the job queue, results in a higher union wage and a smaller union sector. It does not follow immediately that industry unemployment rises, since both labor demand and labor supply contract in this case.

When the reciprocal of the *VMPL* schedule is concave, movement away from a seniority rule *increases* the probability of employment of the median worker. Here two offsetting forces are at work. A majority vote by the original union results in a higher wage, and by itself, a tendency for the least senior workers to leave the sector. But the increased reliance on factors other than seniority in layoff decisions *ceteris paribus* improves prospects for less senior workers, which causes the union to expand, and therefore voting to take place in an environment of less senior workers on average. The outcome depends on the relative strengths of the various forces at work.

## IV. Concluding Remarks

The model developed in this paper takes as its starting point that layoff decisions are based, at least to some extent, on seniority. Also taken as given is the internal political process of the union, which is assumed to be one of majority voting without side payments. Using these assumptions, and information about the stochastic nature of the value marginal product of labor in the unionized sector, the model simultaneously determines the steady-state size of the union sector, and the monopoly-union wage demand. The industry unemployment rate in every state of nature is a by-product.

In order to highlight the roles of the seniority system and the majority voting rule, some important aspects of union wage determination were ignored. First, firms in the unionized sector were assumed to respond passively to union wage demands. A more complete treatment would take into account the role of the bargaining process itself, including the possibility that contracts which Pareto dominate the ones discussed here might emerge. Second, no recognition was taken of the potential heterogeneity of workers. The identity of laid-off workers might be assumed to depend in a predictable way on skill levels. The internal political process in the union becomes more complicated when the union must represent workers with different skill levels, as well as different seniority levels, if the two are not perfectly correlated. Finally, in this paper I have summarily ruled out the possibility that wages vary with seniority or are state-contingent. In general, all union members might prefer these more complicated contracts, if nothing had to be given up in return. Future work might address when such contracts are likely to arise.

Another useful extension of the present paper would make endogenous the rules that govern the ordering of workers in the layoff queue. Here I have simply called this criterion "seniority," and have taken it to be exogenous. By doing so, I have generated a model in which the marginal and median worker differ in a particular way in their preferences over union wage demands. However, if the union's demands for the layoff criterion and the wage rate were to be determined simultaneously, then it is likely that the factors which influence its choice of the former will interact with those that influence its choice of the latter. The nature of this interaction is a subject for future research.

## REFERENCES

**Abowd, John M. and Farber, Henry S.,** "Job Queues and the Union Status of Workers," *Industrial and Labor Relations Review*, April 1982, *35*, 354–67.

**Atherton, Wallace N.,** *Theory of Union Bargaining Goals*, Princeton: Princeton University Press, 1973.

**Azariadis, Costas,** "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

**Blair, Douglas H. and Crawford, David L.,** "Labor Union Objectives and Collective Bargaining," CARESS Working Paper No. 80–24, University of Pennsylvania, 1981.

**Bloch, Farell E. and Kuskin, Mark S.,** "Wage Determination in the Union and Nonunion Sectors," *Industrial and Labor Relations Review*, January 1978, *31*, 183–92.

**Farber, Henry S.,** "Individual Preferences and Union Wage Determination: The Case of the United Mine Workers," *Journal of Political Economy*, October 1978, *86*, 923–42.

**Freeman, Richard B.,** "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, October 1980, *34*, 3–23.

**Grossman, Sanford J. and Hart, Oliver D.,** "Implicit Contracts Under Asymmetric Information," *Quarterly Journal of Economics*, forthcoming.

**Hall, Robert E. and Lilien, David M.,** "Efficient Wage Bargains under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

**Holmstrom, Bengt,** "Equilibrium Long-Term Labor Contracts," *Quarterly Journal of Economics*, forthcoming.

**McDonald, Ian M. and Solow, Robert M.,** "Wage Bargaining and Employment," *American Economic Review*, December 1981, *71*, 968–08.

**Medoff, James L.,** "Layoffs and Alternatives under Trade Unions in U.S. Manufacturing," *American Economic Review*, June 1979, *69*, 380–95.

**_____ and Abraham, Katharine G.,** "Involuntary Terminations Under Explicit and Implicit Employment Contracts," Working Paper No. 634, National Bureau of Economic Research, 1981.

**Weitzman, Martin L.,** "Towards a Theory of Contract Types," Department of Economics Working Paper No. 281, Massachusetts Institute of Technology, 1981.

# Preliminary Evidence on the Use of Inputs by the Federal Reserve System

*By* WILLIAM F. SHUGHART II AND ROBERT D. TOLLISON*

Because the main objectives of the Federal Reserve System (the Fed) concern what Milton Friedman (p. 100) describes as the Holy Trinity—full employment, economic growth, and stable prices—scholarly evaluations of the central bank's performance understandably center most often on the behavior over time of macroeconomic variables such as the aggregate price level, interest rates, and national income. Recently, however, Friedman and others have been drawn to the view that an explanation for the Fed's monetary policy record may be found in the bureaucratic incentives faced by central bank officials.

To illustrate this approach, Friedman observes that the Fed's historical concern with the number of member banks in the system cannot be rationalized on the basis of monetary control; "it derives rather from the prestige, sense of importance, power, and effective lobby that the system gains by supervising many thousands of banks and other financial institutions..." (p. 115). Mark Toma argues that the Fed's independence of the normal congressional budgetary process introduces an inflationary bias to monetary policy. Specifically, he finds a positive relationship between the Fed's income from open market operations and the size of Fed operating expenses.

Keith Acheson and John Chant adduce evidence that several of the Bank of Canada's policy initiatives during the 1950's and 1960's were more consistent with minimizing the cost of government debt than with the Banks' publicly stated goal of extending Canadian financial markets. Chant and Acheson also produce data to support their thesis that the Bank of Canada's "...decisions to emphasize covert instruments, to acquiesce in preserving a concentrated banking system and to maintain an excessive battery of tools appear rational" (1972, p. 30) only when viewed in the context of bureaucratic behavior. Finally, they (1973) find the public pronouncements of the Bank of Canada to be consistent with the bureaucratic goal of avoiding accountability for policy failures.

In this paper we present evidence on Fed behavior that focuses on its use of inputs. Using data that cover the entire period over which the Fed has been in existence, we find a positive and statistically significant *ceteris paribus* relation between the monetary base and the size of the Fed. "Size" is measured in terms of the number of employees. This result suggests that one motivation for expansions in the money supply is to finance the growth of the Fed's bureaucracy. That is, the Fed, which operates under a requirement that it return excess income to the Treasury, "pads" its operating expenditures by increasing the number of employees on its payroll. The employment effect does not explain all (or even most) of historical money supply values. Our empirical results indicate, rather, that bureaucratic incentives play a role in the formulation of monetary policy and perhaps explain in part the Fed's apparent inflationary bias.

Our presentation is organized as follows. Section I contains a discussion of the constraints placed on the Fed's income-earning abilities and indicates how the Fed adjusts to those constraints. In particular, we argue that because the Fed can retain profits denominated in people but not in dollars, it has an incentive to hire more employees than would a for-profit central bank not operating under such a constraint. The basic data and empirical evidence are presented in Section II. Additional findings and implications are reported in Section III, and concluding remarks are offered in Section IV.

---

*Federal Trade Commission and Clemson University, respectively. The opinions expressed in this paper are solely our own and do not necessarily reflect the views of the Federal Trade Commission, its staff, or any individual Commissioner. We benefited from comments on earlier drafts by Phillip Cagan, Robert Ekelund, Richard Higgins, Michael Maloney, James C. Miller III, Dennis Mueller, David Saurman, Gordon Tullock, Myles Wallace, and Thomas Willett. The comments of Robert McCormick and an anonymous referee were especially helpful in improving the paper. The usual caveat applies.

## I. The Fed as a Bureaucratic Agent

As Friedman has suggested, the Fed operates much like any other bureaucracy. As such it seems useful to treat the central bank within a principal-agent context where the Fed carries out a menu of activities for a sponsor. For example, the Fed may pursue the Holy Trinity, it may generate the rate of monetary growth that maximizes inflationary tax revenues, it may be in the business of disrupting private contracts denominated in nominal dollars, it may underwrite budget deficits, and so forth.[1] Whatever functions it performs, and without excluding the possibility that the central bank also serves as agent for the executive branch, we assume that the Fed's sponsor is the Congress and that the Fed's activities are geared toward assisting the legislature in attaining its political goals. The latter may be wealth transfers or macroeconomic stability. In any case, the actual objectives of government are not crucial to our theory of bureaucratic control.

Operationally, the Fed produces and then sells money through commercial banks and a select group of government securities dealers. In carrying out such open market transactions, the Fed operates under several constraints, one of which is that it may purchase only a restricted set of assets. Allowable purchases include U.S. government securities (Treasury bills, notes, certificates, and U.S. bonds) and certain kinds of commercial paper.[2] This securities portfolio represents the principal asset on the central bank's balance sheet; major liabilities include outstanding Federal Reserve notes and member bank reserves.[3]

In an open market purchase of securities, the Fed exchanges the Federal Reserve notes that it has produced at very low resource cost for assets that generate a stream of interest payments.[4] In contrast, open market sales involve the trade of income-earning securities for non-interest-bearing currency. Thus, securities purchases raise the Fed's profits while open market sales reduce its profits.

The Fed pays its current operating expenses out of current interest income, and then returns any excess revenue to the Treasury. The Fed's operating expenses, which are quite similar to those of a typical commercial bank, include maintenance and depreciation of the premises and equipment owned by the Federal Reserve district and branch banks, as well as the offices of the Board of Governors; the costs of operating an extensive check-clearing network; and member bank supervision costs. In most years, the funds turned over to the Treasury are a large multiple of the Fed's current expenses.[5]

How well the Fed assists the legislature in achieving its political goals depends in part on the ability of Congress to monitor its agent. In a bureaucratic setting, economic theory suggests that the degree to which principals can exercise control over the behavior of agents is a function of such factors as the real prices of technical inputs, including accounting and computer services, and political variables such as the size of congressional staffs or the number of committees overseeing the bureau. Of particular importance is that the cost of monitoring any given bureau will depend on the overall number of bureaucratic agents which the principal seeks to control. In the case of Congress, we expect that the secular growth of government, by increasing the number of political margins controlled by the legislature, has raised the costs of monitoring individual bureaus substantially. Moreover, the effect on monitoring costs of increased numbers of depart-

---

[1] Gerald Dwyer has recently produced evidence, however, that "... there is no immediately apparent connection between the federal government's deficits and debt purchases by the Federal Reserve" (p. 327).

[2] See Boris Pesek and Thomas Saving, pp. 102–04.

[3] Federal Reserve notes are not a true liability in the sense that they do not represent a call on the real assets of the Federal Reserve System.

[4] Pesek and Saving observe that "even a moron who gets the right to print government currency (Federal Reserve notes) and sell it for income-earning assets... is bound to show a profit" (p. 407).

[5] See any *Annual Report* of the Board of Governors.

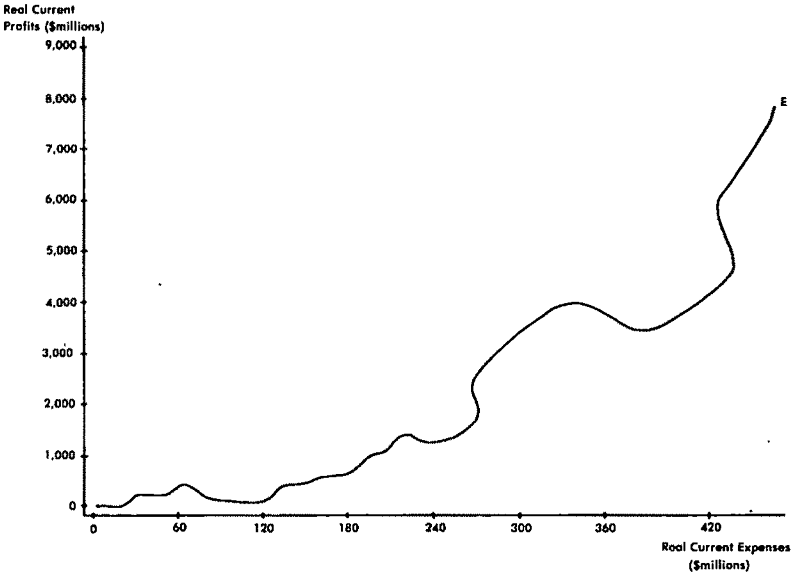


FIGURE 1. THE FED'S PROFITS-AMENITIES TRADEOFF

ments, bureaus, and agencies is likely to swamp any cost-reducing effects arising from lower technical input prices or from improved technologies for reaching consensus among monitors. In this context, it is useful to remember that as government has grown in size over the modern period, the number of legislators/monitors has remained virtually fixed.

As we noted earlier, the Fed's excess income is turned back to the budget. Such a requirement facilitates political equilibrium. If, for example, the Fed does not generate the appropriate level or pattern of transfers, including payments to the Treasury, then at least one political margin will fail to clear: a particular group's tax bill will be "too high"; the gain realized by certain parties to private contracts (debtors, for example) will be "too large"; a government project will receive "too little" funding; and so forth. In that event, some congressmen will be revealed to be inefficient monitors of the Fed and will be replaced with new monitors who are able to constrain the central bank to the equilibrium level or pattern of transfers.

The policing problem with respect to the Fed is not qualitatively different from that of any other principal-agent relation, except that the "bottom line" is political survivorship rather than profits. In the typical analysis of agency problems, it is difficult to obtain empirical surrogates for the relevant variables. Because of the Fed's profit constraint, however, we can formulate a simple agency cost model in which variations over time in monitoring costs affect, and are revealed in, the central bank's use of inputs.

The tradeoff faced by the Fed in allocating its gross income is illustrated in Figure 1. We assume that the Board of Governors maximizes a utility function which includes as arguments the profits earned from open market operations and amenities appearing on the Fed's income statement as operating expenses.[6] The latter includes goods such as salaries, office space, funds for employee travel, and other perquisites. In addition, salaries of the Governors and other officers will depend to some extent on the number of workers they supervise. For simplicity, we assume that the dollar value of amenities is a monotone transformation of Federal Reserve System employment.

Indifference curves labeled $I_0$ and $I_1$ show the expense preferences of the Fed's managers, that is, the various combinations of amenities and profits that yield a given level of utility. The rate at which profits can be transformed into amenities in the market is given by the budget lines $B_0$ and $B_1$.

---

[6] We adopt here George Stigler's (1966, p. 73) formulation of the pioneering work of Armen Alchian and Reuben Kessel.

The opportunity set, $B_0$, represents the constraint that would prevail if the Congress did not require the Fed to return excess income to the Treasury. In that case, equilibrium would obtain at a point such as $P$ where an indifference curve, $I_0$, is just tangent to $B_0$. Here, the Fed earns profits in the amount $0E$, and purchases $0C$ amenities with foregone profits, $AE$.

Now suppose that Congress designates the Treasury as residual claimant of the Fed's excess income. To the extent that such a constraint is effective, the budget line will become less steep, indicating that the Fed now faces smaller costs for amenities in terms of foregone profits. In essence, the constraint allows the Federal Reserve System to retain profits that are denominated in employees or other amenities, but not those denominated in dollars. A new equilibrium consequently occurs at $P'$ where $I_1$ is tangent to $B_1$. The Fed now foregoes $FE'$ in dollar profits and purchases $0C'$ amenities. Funds in the amount $0E'$ are turned over to the government's general revenue account.[7] Comparing this solution to the equilibrium with no constraint on how profits may be allocated, it is plain that the Fed buys more noble premises, hires more employees, and takes advantage of more perquisites than before.

The comparative statics result depicted in the diagram of course depends on the assumption that amenities are a normal good. In that case, a fall in the price of amenities in terms of dollar profits will lead the Fed to purchase more of them if the substitution effect dominates the income effect. In general, however, the overall effect on the quantity of amenities is indeterminate a priori.

Congress certainly recognizes that the profit constraint induces the Fed to overpurchase amenities, but it will generally not pay to eliminate entirely such effects if the opportunity cost of monitoring the central bank is positive. As we noted before, as the size of government grows, a fixed number of legislators/monitors must monitor more political margins, and the cost of policing any individual agency consequently rises.

Expansionary open market operations enlarge the Fed's opportunity set, that is, shift $B_1$ to the right, raising potential profits (and therefore potential amenities) for the central bank. If the relative prices of profits and amenities remain constant over time, each successive shift in $B_1$ will be parallel to the previous budget line. In that case, the Fed's "expansion" path would be given by ray 2, which lies to the right of the path that would be followed by a bank whose choices were constrained only by market prices (ray 1). On the other hand, rising congressional monitoring costs will cause counterclockwise rotations in the constraint as the Fed's opportunity set moves to the right. These successively flatter budget lines will yield an expansion path for amenities, such as ray 3, which lies to the right of and below the path that would obtain if relative prices remain constant. That is, increasingly higher policing costs resulting from government growth will generate a clear "amenities bias" in the Fed's input purchases.

Figure 1 illustrates two general consequences of requiring the Fed to turn back its excess income to the Treasury. In a timeless setting, the constraint changes the relative prices of amenities and profits, leading the Fed to purchase more amenities if the substitution effect dominates the income effect. Over time, rising monitoring costs cause the Fed's input choices to diverge increasingly from the path that would be chosen by a central bank not operating under the postulated conditions.

---

[7] The diagram is not intended to show the actual relative magnitudes of foregone profits and operating expenses. During 1981, the twelve Federal Reserve Banks earned \$15.51 billion in income, \$14.55 billion or 94 percent of which was attributable to the Fed's portfolio of U.S. government securities. After paying their operating expenses, contributing \$63.16 million to defray the expenditures of the Board of Governors, and making other accounting adjustments, the banks turned over \$14.02 billion in profits to the government's general revenue account as "interest on Federal Reserve notes." For the entire 1915–81 period, the Fed retained on average 14 percent of its income, contributing \$98.79 billion to the Treasury out of \$115.10 billion in total revenues.

As expansionary monetary policies enlarge the Fed's potential profits-amenities choices, the Fed faces higher real costs of operating the monetary system. That is, expansions in the money supply undoubtedly require additional inputs for the purposes of, for example, clearing a larger number of checks, engaging in a larger number of open market transactions, and so forth. Under a fiat standard, however, these incremental production costs will be trivial in comparison to the nominal value added to the money stock. Our model predicts in addition that the Fed will tend to purchase more amenities than would be justified by cost-minimizing money production, that is, to incur "fake" costs.

That the Fed does not expend all of its income, but chooses to return some profits to the Treasury, implies a survivorship strategy. Our purpose here is not to explain the political constraints that guide such strategies, but rather to examine Fed behavior under the postulated conditions. Moreover, it is important to emphasize that the tendency of the Fed to purchase more amenities than would a for-profit producer of money is not due to the "waste" or lack of efficiency sometimes attributed to monopoly, but is in addition to such arguments.[8] Rather, the Fed chooses a combination of amenities and payments to the Treasury that maximizes its utility subject to the profit-confiscating political constraint.

The Fed's incentive to enlarge its purchases of amenities conflicts with its mandate to promote macroeconomic stability only part of the time. When engaging in expansionary policies, the Fed can both increase the supply of money and increase the size of its bureaucracy because the two goals are served by open market purchases of securities. In contrast, contractionary policies force the Fed to lower its profits when it undertakes open market sales to reduce the rate of growth in money. It is an empirical question whether the Fed's concern with the size of its bureaucracy is useful in explaining the way in which monetary policy is carried out. It is to that question that we now turn.

## II. Empirical Evidence

We have outlined a hypothesis which suggests that the Federal Reserve System may operate like a typical bureaucracy. Specifically, because the Fed must return excess revenues to the Treasury, it has an incentive to take a portion of its profits in the form of amenities rather than dollars. Since the success of bureaucratic organizations is often measured by the relative size of staffs, we expect the empirical importance of our hypothesis to be revealed in terms of the number of officers and employees in the Federal Reserve System, although the Fed's propensity to enlarge its purchases of amenities could take the form of more sumptuous offices, larger salaries, or other perquisites.[9]

### A. *Regression Model*

We estimated the parameters of the following money supply equation to test the predictions of our hypothesis about amenities and profits:

$$\begin{aligned} BASE = b_0 + b_1\, SYSEMP + b_2\, CD \\ + b_3\, INT + + b_4\, RGNP \\ + b_5\, TREND + b_6\, TREND^2 + e. \end{aligned}$$

The variables are defined as follows: *BASE* is the nominal monetary base; *SYSEMP* is the number of Federal Reserve System employees; *CD* is the actual currency-deposit ratio; *INT* is the interest rate on prime commercial paper of four- to six-months' maturity; *RGNP* is real Gross National Product; *TREND* is a linear time trend; and *e* is the regression error term.

The principal prediction of our hypothesis is that the sign on *SYSEMP* will be positive. As we noted earlier, the Fed produces money and sells it to commercial banks and a small group of securities dealers in exchange for income-earning assets. Although the definition of money is a much-debated concept, there are two principal types in the U.S. economy: currency sold by the Fed (Federal Reserve notes) and money produced by commercial banks in the form of demand deposits. The stock of nominal money, referred to as $M_1$, is the sum of currency in the hands of the public and the dollar value of demand deposits outstanding.[10]

---

[8] On this point, see Stigler (1976).

[9] Gordon Tullock informs us that the "old" Federal Reserve Board Building was for many years the building in Washington with the highest cost per square foot.

The portion of the money supply produced by the Fed consists of currency in the hands of the public and bank reserves. Fed money is referred to as the "monetary base" or "high-powered money." In its simplest formulation the relation between $M_1$ and the monetary base is an identity that depends on two ratios. These are the currency-deposit ratio which gives the proportion in which the public holds the components of $M_1$, and the reserve-deposit ratio which measures the amount of currency held by banks against their outstanding demand deposit liabilities.

For given values of the two deposit ratios, increases in the stock of high-powered money raise $M_1$. Expansionary policies such as open market purchases of securities both raise the monetary base and increase the Fed's profits. Since higher profits can be used by the Fed to purchase more amenities, the sign on *SYSEMP* is predicted to be positive, *ceteris paribus.*

Our regression equation in other respects is a traditional money supply function. Under the typical assumption that the Fed actively pursues a stabilization or control function, the monetary authority will use macroeconomic data to guide it in deciding whether to expand or contract the money supply by changing the stock of high-powered money. Three monetary indicators suggested in the literature are the actual currency-deposit ratio, interest rates, and the level of real income.

The sign on *CD* is predicted to be positive. Other things equal, the Fed responds to increases in the public's holdings of currency relative to demand deposits by expanding the monetary base. It does so because increases in the currency-deposit ratio lead to reductions in bank reserves. Since reserves are the basis on which bank money is supplied, the lower level of reserves causes a contraction in $M_1$. This effect can be offset by increasing the supply of high-powered money. Thus, if the Fed uses a countercyclical feedback rule in which a higher currency-deposit ratio signals the need for expansionary policies, the sign on *CD* will be positive.

Until quite recently, the Fed used interest rates as a monetary control target. *Ceteris paribus*, higher interest rates led the monetary authority to expand the base in order to increase the supply of available credit. Similarly, a contractionary policy was called for whenever the rate of interest reached relatively low levels. Other things equal, the regression coefficient on *INT* should therefore be positive.

Increases in the level of real income will lead to expansions in the monetary base if the Fed supplies the amount of real balances desired by the public. Other things equal, an increase in income raises the demand for money because at higher income levels individuals hold larger money balances to finance a greater number of market transactions. Moreover, raising the stock of high-powered money when income increases will counteract the tendency for prices to fall. The sign on *RGNP* should therefore be positive.

Finally, a time trend is included to account for secular growth in Fed employment. We might expect the number of Federal Reserve System employees to rise over time as part of the increase in the size of government as a whole. Such tendencies, which are exogenous to the profits-amenities relation, will be captured in the coefficient on *TREND*, whose sign is likely to be positive. Squaring the trend variable allows for nonlinear growth over time.

### B. *The Data*

We estimated the regression model using data from the Federal Reserve System's *Annual Report* and from other sources. (The basic data are shown in Table 1.) We define system employment to include the officers and employees of the district and branch

---

[10] The Fed supplies the nominal base; the real stock of money is determined by the public.

TABLE 1—FEDERAL RESERVE EMPLOYMENT, MONETARY AGGREGATES, *GNP*, AND COMMERCIAL PAPER RATE

| Year | Paid Civ. Empl. of Federal Govt. | Empl. of Board of Governors | Empl. of Fed. Res. Banks | Currency ($bil.) | Demand Deposits ($bil.) | Bank Reserves ($bil.) | Monetary Base ($bil.) | *GNP* ($bil.) | Prime Commercial Paper Rate (Percent) |
|---|---|---|---|---|---|---|---|---|---|
| 1915 | 395429 | 59 | 442 | 1.93 | 10.55 | 0.45 | 2.38 | 40.0 | 4.01 |
| 1916 | 399381 | 61 | 920 | 2.17 | 12.53 | 0.88 | 3.05 | 48.3 | 3.84 |
| 1917 | 438500 | 77 | 2985 | 2.17 | 14.91 | 1.58 | 3.75 | 60.4 | 5.07 |
| 1918 | 834500 | 227 | 7706[a] | 2.76 | 16.20 | 1.81 | 4.57 | 76.4 | 6.02 |
| 1919 | 794271 | 319[c] | 9459[b] | 4.02 | 17.77 | 2.02 | 6.04 | 84.0 | 5.37 |
| 1920 | 655265 | 255[c] | 12285[d] | 4.48 | 19.25 | 1.86 | 6.34 | 91.5 | 7.50 |
| 1921 | 561142 | 249[e] | 12745 | 4.04 | 17.47 | 1.88 | 5.92 | 69.6 | 6.62 |
| 1922 | 543507 | 211[c] | 12256[f] | 3.69 | 17.98 | 1.97 | 5.66 | 74.1 | 4.52 |
| 1923 | 536900 | 209 | 11996[g] | 3.96 | 18.98 | 1.96 | 5.92 | 85.1 | 5.07 |
| 1924 | 543484 | 197 | 10761[g] | 3.96 | 19.71 | 2.31 | 6.27 | 84.7 | 3.98 |
| 1925 | 553045 | 181 | 10126[g] | 3.96 | 21.70 | 2.26 | 6.22 | 93.1 | 4.02 |
| 1926 | 548713 | 203 | 10280[g] | 4.00 | 22.18 | 2.28 | 6.28 | 97.0 | 4.34 |
| 1927 | 547127 | 201 | 10108 | 3.98 | 22.12 | 2.53 | 6.51 | 94.9 | 4.11 |
| 1928 | 560772 | 196 | 9989[h] | 3.89 | 22.49 | 2.44 | 6.33 | 97.0 | 4.85 |
| 1929 | 579559 | 234 | 9986[h] | 3.90 | 22.74 | 2.41 | 6.31 | 103.1 | 5.85 |
| 1930 | 601319 | 209 | 9609[i] | 3.73 | 22.03 | 2.52 | 6.25 | 90.4 | 3.59 |
| 1931 | 609746 | 197 | 9340[j] | 4.16 | 19.98 | 2.13 | 6.29 | 75.8 | 2.64 |
| 1932 | 605496 | 203 | 10319[k] | 4.92 | 16.19 | 2.56 | 7.48 | 58.0 | 2.73 |
| 1933 | 603587 | 261 | 12042[l] | 5.09 | 14.82 | 2.87 | 7.96 | 55.6 | 1.73 |
| 1934 | 698649 | 320 | 11706 | 4.63 | 17.23 | 4.41 | 9.04 | 65.1 | 1.02 |
| 1935 | 780582 | 334 | 11551 | 4.80 | 21.08 | 6.39 | 11.19 | 72.2 | 0.75 |
| 1936 | 867432 | 323 | 10390 | 5.23 | 24.32 | 7.11 | 12.34 | 82.5 | 0.75 |
| 1937 | 895993 | 401 | 10385 | 5.59 | 25.32 | 7.58 | 13.17 | 90.4 | 0.94 |
| 1938 | 882226 | 416 | 10919 | 5.55 | 24.97 | 10.09 | 15.64 | 84.7 | 0.81 |
| 1939 | 953891 | 426[m] | 11355 | 6.04 | 28.11 | 11.47 | 17.51 | 90.9 | 0.59 |
| 1940 | 1042420 | 435[m] | 11640 | 6.76 | 32.89 | 14.07 | 20.83 | 100.0 | 0.56 |
| 1941 | 1437682 | 455 | 14083 | 8.40 | 38.12 | 12.81 | 21.21 | 125.0 | 0.53 |
| 1942 | 2296384 | 422 | 19972 | 11.54 | 43.82 | 13.15 | 24.69 | 158.5 | 0.66 |
| 1943 | 3299414 | 459 | 24741 | 16.35 | 55.89 | 12.75 | 29.10 | 192.1 | 0.69 |
| 1944 | 3332356 | 448 | 24442 | 21.22 | 64.12 | 14.17 | 35.39 | 210.6 | 0.73 |
| 1945 | 3816310 | 455 | 24442 | 25.33 | 73.90 | 16.03 | 41.36 | 212.4 | 0.75 |
| 1946 | 2696529 | 480 | 21430 | 26.48 | 79.98 | 16.52 | 43.00 | 209.8 | 0.81 |
| 1947 | 2111001 | 504 | 19364 | 26.58 | 85.22 | 17.26 | 43.84 | 233.1 | 1.03 |
| 1948 | 2071009 | 527 | 19081 | 26.07 | 86.24 | 19.99 | 46.06 | 259.5 | 1.44 |
| 1949 | 2102109 | 548 | 17967 | 25.50 | 85.67 | 16.29 | 41.79 | 258.3 | 1.49 |
| 1950 | 1960708 | 578 | 17993 | 25.05 | 89.08 | 17.39 | 42.44 | 286.5 | 1.45 |
| 1951 | 2482666 | 586 | 19832 | 25.53 | 93.67 | 20.31 | 45.84 | 330.8 | 2.16 |
| 1952 | 2600612 | 568 | 20011 | 26.70 | 98.52 | 21.18 | 47.88 | 348.0 | 2.33 |
| 1953 | 2558416 | 581 | 20272 | 27.78 | 100.64 | 19.92 | 47.70 | 366.8 | 2.52 |
| 1954 | 2407676 | 558 | 19334 | 27.52 | 102.75 | 19.28 | 46.80 | 366.8 | 1.58 |
| 1955 | 2397309 | 559 | 19134 | 27.63 | 106.79 | 19.24 | 46.87 | 400.0 | 2.18 |
| 1956 | 2398736 | 586 | 19307 | 27.98 | 108.05 | 19.54 | 47.52 | 421.7 | 3.31 |
| 1957 | 2417565 | 580 | 19817 | 28.26 | 108.48 | 19.42 | 47.68 | 444.0 | 3.81 |
| 1958 | 2382491 | 603 | 19307 | 28.37 | 109.98 | 18.90 | 47.27 | 449.7 | 2.46 |
| 1959 | 2382807 | 600 | 19500 | 28.90 | 114.38 | 18.93 | 47.83 | 487.9 | 3.97 |
| 1960 | 2398704 | 596 | 19931 | 28.99 | 112.62 | 19.28 | 48.27 | 506.5 | 3.85 |
| 1961 | 2435804 | 599 | 20023 | 29.10 | 114.82 | 20.12 | 49.22 | 524.6 | 2.97 |
| 1962 | 2514197 | 601 | 20380 | 30.09 | 116.91 | 20.04 | 50.13 | 565.0 | 3.26 |
| 1963 | 2527960 | 615 | 20269 | 31.55 | 119.74 | 20.75 | 52.30 | 596.7 | 3.55 |
| 1964 | 2500503 | 632 | 19468 | 33.49 | 123.74 | 21.61 | 55.10 | 637.7 | 3.97 |
| 1965 | 2527915 | 648 | 18687 | 35.26 | 128.54 | 22.72 | 57.98 | 691.1 | 4.38 |
| 1966 | 2759019 | 665 | 19131 | 37.48 | 133.58 | 23.83 | 61.31 | 756.0 | 5.55 |
| 1967 | 3002461 | 753 | 18996 | 39.37 | 138.38 | 25.26 | 64.63 | 799.6 | 5.10 |
| 1968 | 3055212 | 796 | 19607 | 41.97 | 148.47 | 27.22 | 69.19 | 873.4 | 5.90 |
| 1969 | 3076414 | 902 | 20755 | 44.82 | 156.94 | 28.03 | 72.85 | 944.0 | 7.83 |
| 1970 | 2921909 | 1047 | 22079 | 47.69 | 162.30 | 29.27 | 76.96 | 992.7 | 7.72 |

TABLE 1—*(Continued)*

| Year | Paid Civ. Empl. of Federal Govt. | Empl. of Board of Governors | Empl. of Fed. Res. Banks | Currency ($bil.) | Demand Deposits ($bil.) | Bank Reserves ($bil.) | Monetary Base ($bil.) | *GNP* ($bil.) | Prime Commercial Paper Rate (Percent) |
|---|---|---|---|---|---|---|---|---|---|
| 1971 | 2862894 | 1169 | 22480 | 52.60 | 176.80 | 31.33 | 83.93 | 1077.6 | 5.11 |
| 1972 | 2811779 | 1208 | 24254 | 56.90 | 193.60 | 31.35 | 88.25 | 1185.9 | 4.73 |
| 1973 | 2765664 | 1276 | 26655 | 61.60 | 202.50 | 35.07 | 96.67 | 1326.4 | 8.15 |
| 1974 | 2893119 | 1363 | 27637 | 67.80 | 207.40 | 36.94 | 104.74 | 1434.2 | 9.84 |
| 1975 | 2896944 | 1445 | 26515 | 73.80 | 214.10 | 34.99 | 108.79 | 1549.2 | 6.32 |
| 1976 | 2883134 | 1470 | 25272 | 80.60 | 224.40 | 35.14 | 115.74 | 1718.0 | 5.34 |
| 1977 | 2841152 | 1477 | 24114 | 88.60 | 239.70 | 36.47 | 125.07 | 1918.0 | 5.61 |
| 1978 | 2872851 | 1477 | 23373 | 97.40 | 253.90 | 41.57 | 138.97 | 2156.1 | 7.99 |
| 1979 | 2875872 | 1451 | 23416 | 106.10 | 262.80 | 43.97 | 150.07 | 2413.9 | 10.91 |
| 1980 | 2875866 | 1525 | 24208 | 116.10 | 267.40 | 40.10 | 156.20 | 2626.1 | 12.29 |
| 1981 | 2895712 | 1504 | 23696 | 123.10 | 236.80 | 42.01 | 165.11 | 2922.2 | 14.76 |

*Sources:* Paid civilian employees of the federal government, 1915–64, from U.S. Department of Commerce, *Historical Statistics*. The years 1965–81 were taken from U.S. Department of Commerce, *Statistical Abstract*, various years. Employees of Federal Reserve System are listed in Board of Governors of the Federal Reserve System, *Annual Report*, various years. Currency in the hands of the public, demand deposits, and bank reserves, 1915–70, are from *Historical Statistics*; data for 1971–81 are from the *Economic Report of the President*. Gross National Product, 1915–39, is given in *Historical Statistics*; data for the years 1940–81 are contained in the *Economic Report of the President*. The implicit *GNP* price deflator was spliced to change the base year from 1958 to 1972. The interest rate on prime commercial paper of four- to six-months maturity was taken from *Historical Statistics* for the years 1915–70, and from the *Economic Report of the President* for 1971–81.

[a] Includes 6 part-time employees.

[b] Includes 111 detailed currency counters.

[c] As of December.

[d] Includes 34 detailed employees and 2 part-time employees.

[e] Includes 19 detailed employees, as of December.

[f] Includes 3 detailed employees and 1 part-time employee.

[g] Includes 2 detailed employees.

[h] Includes 9 part-time employees.

[i] Includes 10 part-time employees.

[j] Excludes 15 part-time employees paid on an hourly basis.

[k] Excludes 14 part-time employees paid on an hourly basis.

[l] Excludes 14 part-time employees and 4 temporary employees paid on an hourly basis, and 3 special temporary employees paid on a monthly basis.

[m] Estimated.

banks as well as those of the Board of Governors. The observations are on an annual basis and cover the period from 1915 through 1981. Although the Federal Reserve Act was passed in late 1913, the Fed did not officially open its doors for business until the fall of 1914; 1915 was the first full year of operation for the central bank.

All variables, except for the interest rate and time trend, were transformed by taking the natural logarithm of the observations. Initial estimates using ordinary least squares indicated significant autocorrelation in the residuals. The equation was therefore reestimated using generalized least squares.[11] We also allowed for feedback effects on monetary policy by using the values of the currency-deposit ratio and interest rate lagged one year.

---

[11] A. C. Harvey (pp. 189–92) discusses the relative small-sample efficiency of generalized least squares to the Cochrane-Orcutt method when the model contains a trending explanatory variable.

TABLE 2—REGRESSION RESULTS

| | | | | |
|---|---|---|---|---|
| Constant | −1.0233 | −1.0863 | −1.5327 | −1.2860 |
| *LNSEMP* | 0.1308 | 0.1503 | 0.1523 | 0.1497 |
| | (2.66)*** | (3.63)*** | (3.03)*** | (3.22)*** |
| *LNCD* | 0.3357 | 0.2703 | | |
| | (3.14)*** | (2.73)*** | | |
| *LNCD*(−1) | | | 0.3437 | 0.2774 |
| | | | (2.94)*** | (2.54)** |
| *INT*(−1) | −0.0068 | | −0.0052 | |
| | (−0.75) | | (−0.54) | |
| *LNRGNP* | 0.3108 | 0.2615 | 0.3757 | 0.3079 |
| | (1.95)* | (1.79)* | (2.19)** | (1.96)* |
| *TREND* | 0.0569 | 0.0596 | 0.0537 | 0.0567 |
| | (6.13)*** | (6.38)*** | (5.71)*** | (6.12)*** |
| $TREND^2$ | −0.0002 | −0.0002 | −0.0002 | −0.0002 |
| | (−1.80)* | (−2.07)** | (−1.74)* | (−2.02)** |
| $\hat{\rho}$ | 0.7657 | 0.8234 | 0.7329 | 0.7939 |
| $R^2$ | 0.952 | 0.940 | 0.956 | 0.943 |

*Notes:* Dependent Variable: Natural Logarithm of Nominal Monetary Base, 1915–81. $t$-values are given in parentheses; $\hat{\rho}$ is the estimated autocorrelation coefficient. Asterisks denote significance at the 1 percent (***), 5 percent (**), and 10 percent (*) levels.

### C. *The Results*

The regression results are given in Table 2. The estimated coefficients in column 1 provide broad confirmation of the profits-amenities hypothesis. An increase in the stock of high-powered money is strongly associated with an increase in the number of employees on the Fed's payroll. The coefficient is statistically different from zero at the 1 percent level. Moreover, the contemporaneous currency-deposit ratio has the expected sign and is equally significant. An increase in real income also leads to expansions in the monetary base; the coefficient reaches significance at the 10 percent level.

The insignificant coefficient on the interest rate variable conflicts with our a priori expectations. It is possible that this result obtains because we have used nominal rather than real interest rates, or because money demand effects dominate the postulated supply relationship. In addition to its role as a monetary target, the rate of interest represents the opportunity cost of holding money. Because individuals must forego the income from interest-bearing assets by holding money, other things equal, they reduce their money balances as interest rates rise.

In column 3, the contemporaneous currency-deposit ratio is replaced by the value of the same variable lagged one year. This specification would be appropriate if the Fed uses a feedback rule in which this year's monetary base is affected by last year's currency-deposit ratio. Lagged interest rates are deleted in column 4. The overall regression results seem to be unaffected by this substitution.[12]

The models in Table 2 explain 94 percent or more of the variation in the monetary base. Since we have run the regressions in logs, the coefficients can be directly interpreted as elasticities. The coefficient on the employment variable, *LNSEMP*, is about .14, indicating that a 1 percent increase in Fed employment results in more than a one-tenth of 1 percent rise in the monetary base. Over the 1915–81 period, the mean nominal stock of high-powered money was \$44.7 billion, and there were on average 17,328 workers on the Fed payroll. Thus, each additional 173 Fed employees added roughly \$62.6 million to the monetary base. Put another way, every time the Fed hired one more worker, the stock of high-powered money rose by about \$362,000. Assuming these funds were used to purchase securities yielding a 10 percent return, the \$362,000 would generate a thirty-year annuity of \$20,745. This appears to be about the order of magnitude required to support the career of an average Fed employee.[13]

---

[12] We estimated an alternative specification which involved including the natural logarithm of the implicit price deflator as an explanatory variable. The price variable would be appropriate if the Fed uses a feedback rule in which higher prices lead to contractionary policies, and vice versa. The estimated coefficient on this parameter was not significantly different from zero, however. In addition, results qualitatively similar to those in Table 2 were obtained when percentage changes in the variables were used.

## III. Additional Empirical Findings

### A. *Relative Fed Growth*

We considered whether the growth of the Fed relative to the federal government as a whole would explain variations in the monetary base. Such a formulation would be appropriate if the central bank's behavior is driven in part by a desire for prestige, which Acheson and Chant define as "...the position of the bureau in the social hierarchy of government bureaus" (p. 639). To test this proposition, we replaced *LNSEMP* by the natural logarithm of the ratio of Fed employment to the total number of paid civilian workers in the federal government, *LNSYSPCT*, and obtained the following results:

$$(1)\quad \begin{aligned} LNBASE = -0.4534 + &\underset{(2.42)^{**}}{0.1094 LNSYSPCT} \\ &+ \underset{(3.63)^{***}}{0.3824 LNCD} + \underset{(3.13)^{***}}{0.5003 LNRGNP} \\ &+ \underset{(7.37)^{***}}{0.0657 TREND} - \underset{(-3.80)^{***}}{0.0004 TREND^2} \\ &\hat{\rho} = 0.7908;\ R^2 = 0.944. \end{aligned}$$

The coefficient on the Fed employment ratio is positive and significantly different from zero at the 5 percent level. This lower significance level is undoubtedly due to the tendency of increases in total government employment to swamp increases in the number of workers on the central bank's payroll. However, the overall results are similar to those obtained when the absolute number of Fed employees is used as an explanatory variable.

The result in equation (1) suggests that increases in the base are used to finance increases in the Fed's *relative* bureaucratic standing. That is, the employment ratio coefficient can be interpreted as showing that the Fed has been more successful in enlarging its employee staff over time than the federal government as a whole.

### B. *Fed Growth and the Velocity of Money*

We investigated whether the positive and significant coefficient on the employment variable could be explained by spurious correlation arising because *LNSEMP* was picking up changes in the velocity of money. Velocity measures the number of times each \$1 turns over in financing a given level of nominal transactions. If more Fed employees are required, for example, to clear a larger volume of checks during times when velocity is relatively high, the estimated base-amenities relationship would be misleading.

Using the logarithm of contemporaneous velocity, *LNV*, as an explanatory variable in the regression equation, we obtained the following results:

$$(2)\quad \begin{aligned} LNBASE = 0.2980 + &\underset{(5.19)^{***}}{0.2384 LNSEMP} \\ &+ \underset{(1.41)}{0.1510 LNCD} - \underset{(-3.74)^{***}}{0.6505 LNV} \\ &+ \underset{(3.72)^{***}}{0.0386 TREND} + \underset{(1.58)}{0.0003 TREND^2} \\ &\hat{\rho} = 0.7052;\ R^2 = 0.967. \end{aligned}$$

The estimates are quantitatively similar to those obtained in the other specifications.

---

[13] Employment at the Board of Governors is highly correlated with employment in the district and branch banks ($r^2 = 0.804$, which is significantly different from zero at the 1 percent level). The regression results are therefore unaffected by our choice of System employment as the input measure.

TABLE 3—TEST OF CAUSALITY

| | Lags | | | | | | |
|---|---|---|---|---|---|---|---|
| Constant | −2 | −1 | 0 | +1 | +2 | $\hat{\rho}$ | $R^2$ |
| Monetary Base Regressed on Past and Future Fed Employment | | | | | | | |
| 0.0493 | 0.0606 | 0.1343 | 0.0102 | 0.0492 | −0.1258 | 0.4295 | 0.226 |
| | (1.02) | (1.89)* | (0.14) | (0.41) | (−1.04) | | |
| 0.0481 | 0.0782 | 0.1245 | −0.0029 | | | 0.4284 | 0.211 |
| | (1.39) | (1.81)* | (−0.04) | | | | |
| Fed Employment Regressed on Past and Future Monetary Base | | | | | | | |
| −0.0601 | 0.1942 | 0.3236 | 0.3093 | 0.8119 | 0.0027 | 0.2619 | 0.307 |
| | (0.80) | (1.27) | (1.21) | (3.13)*** | (0.01) | | |
| −0.0359 | 0.3248 | 0.3362 | 0.5209 | | | 0.2042 | 0.192 |
| | (1.26) | (1.20) | (1.93)* | | | | |

*Notes:* See Table 2.

The negative sign on *LNV* is consistent with a countercyclical feedback rule of the following form—other things equal, a rise in velocity lowers the public's demand for money because a higher turnover rate implies that each $1 finances a higher proportion of transactions. Contractions in the monetary base will counteract the tendency for prices to rise as velocity increases.

The results obtained here as well as those in equation (1) suggest that Fed employment has not risen merely because more people are required to supervise larger money stocks. On the contrary, the bureaucratic aspects of Fed behavior appear to be powerful enough to overcome the substantial scale economies in fiat money production.[14]

### C. *Test of Causality*

A third important issue is whether an increase in Fed employment leads to expansions in the monetary base or, alternatively, changes in the stock of high-powered money result in changes in the number of employees. To answer this question, we performed the test of causality suggested by C. W. J. Granger and later formalized by Christopher Sims.

After detrending the data by taking first differences of the logged observations, Sims' test involves estimating two-sided distributed lag regressions. For the hypothesis that there is no feedback from current changes in the monetary base to future Fed employment, one regresses the current monetary base on past, contemporaneous, and future employment figures. For the reverse hypothesis that there is no feedback from current employment levels to future base changes, this year's change in employment is regressed on past and future money.

The results are reported in Table 3. For each dependent variable, one regression includes two future and two past values of the right-hand side variable; the other regression omits the leading lags. In the first row, the small and insignificant coefficients on future changes in employment do not lead one to reject the null hypothesis that money does not cause employment in the sense of Granger. Moreover, comparing the overall results with those in row 2, the relevant $F$-statistic implies that the coefficients on the future values of employment are zero.[15]

---

[14] The results are also consistent with Friedman's (p. 113) contention of bond market "churning." He observes that in 1980 the Fed's $800 billion of open market purchases led to a net change of only $4.5 billion in the bank's securities portfolio. Such an "excessive" number of transactions may result from an excessive number of securities traders.

[15] Let $K$ denote the number of independent variables in the regression equation, which omits the leading values, and $Q$ the number of right-hand-side arguments for the full expression. If the added variables ($Q > K$)

In contrast, the results in rows 3 and 4 are consistent with the hypothesis that employment causes money. The coefficient on the one-year-ahead money value is significantly different from zero and is absolutely larger than the coefficients on the lagged money variables. Moreover, the calculated $F$ is 4.65, which exceeds the critical value at the 5 percent level.

We conclude that the appropriate regression specification is one in which the monetary base is the dependent variable. That is, employment belongs in the money supply equation; the base does not belong in an input demand function. This result provides further support for the notion that bureaucratic incentives drive the money supply process and contradict the alternative view that increases in the base result in more employee hiring for the purpose of handling larger numbers of accounting entries.[16]

### D. *Implications*

In Section I, we noted that Fed employment contains both "real" and "fake" components, the latter representing that portion of the payroll not justified by cost-minimizing money production. The theory also predicted that the postulated budget constraint would lead the Fed to over-purchase amenities (and therefore have a flatter expansion path) relative to a for-profit central bank. The empirical discussion has focused on the strong employment effect in the data without showing that the Fed's profit-amenities choices are different from those of the for-profit benchmark.

We have reported three pieces of evidence which are consistent with our theory. First, the positive base-employment relationship was not weakened by including velocity (transactions per dollar) as an explanatory variable. This result indicated that Fed employment rose by more than the amount required to perform the additional accounting work arising from increased velocity. Second, the causality tests implied that Fed employment properly appears in the money supply function, and not vice versa. People cause money; money does not cause people. Third, we found that growth of central bank employment relative to the federal government as a whole was associated with increases in the base. It is difficult to explain any of these results on the basis of hypotheses other than bureaucratic expansionism, given the very large scale economies in fiat money production.

As a fourth element of support for the bureaucratic hypothesis, consider the plot of the Fed's actual "expansion" path in Figure 2. The amount of real current profits appears on the vertical axis; contemporaneous real expenditures are on the horizontal axis. The data include the central bank's current income from all sources, the bulk of which arises from open market operations, and all current expenses, that is, both real and fake

---

are not relevant in explaining variations in the dependent variable, then differences in the regression sums of squares would be attributable to sampling error. The test statistic has the following form:

$$\left(\frac{R_Q^2 - R_K^2}{1 - R_Q^2}\right)\left(\frac{n-Q}{Q-K}\right) \sim F_{Q-K,\,n-Q},$$

where $n$ is the number of observations. For the first regression pair, $R^2$ rises trivially when the leading employment values are added. The two variables do not therefore contribute significantly to the equation's explanatory value.

[16] We also estimated all regression specifications augmented by a dummy variable that took on the value of unity for the years 1934–81, and zero otherwise. The dummy variable was employed to account for a possible effect connected with repeal in 1933 of the requirement that the Fed pay half of its yearly *net* earnings to the Treasury. Prior to this change, which Toma describes as a quid pro quo for creation of the Federal Deposit Insurance Corporation, the Board of Governors was funded through the normal congressional budgetary process. The member banks paid their own way, however, and because the turnover requirement was expressed in terms of net earnings, it is not entirely clear a priori how this constraint might have affected the base-amenities relation. However, the results invariably showed a positive and significant coefficient on the dummy variable with no change in either the magnitude or significance of other coefficients. This suggests that removal of the half-yearly-net-earnings turnover requirement led to greater monetary expansionism, without affecting the Fed's amenities bias. If anything, the bias was strengthened in the sense that the coefficient on the Fed employment ratio in equation (1) was significant at the 1 percent level when the dummy variable was included.

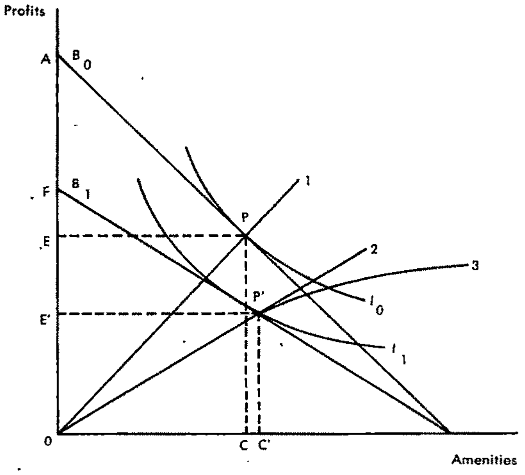


FIGURE 2. THE FED'S ACTUAL "EXPANSION" PATH

costs. In a given year, the sum of the two figures represents the Fed's gross income.

At low values of the two variables, the expansion path is virtually horizontal, indicating that during the early years of its existence the Fed expensed a large proportion of its income. Until the end of the 1930's the Fed turned back roughly one-third to one-half of its annual securities portfolio income to the Treasury. In later years, the expansion path moves slowly upward and to the right, showing a strong amenities bias that is consistent with our theory. The Fed's actual expansion path appears remarkably similar to ray 3 in Figure 1, and is clearly not the type of path one would expect with, for example, substantial economies of scale in fiat money production.

## IV. Concluding Remarks

In this paper we have presented evidence of a positive and statistically significant relationship between changes in the size of the Fed and changes in the nominal stock of high-powered money. This result suggests that the Federal Reserve System may function like a typical bureaucracy. That is, the Fed seems to adjust to the requirement that it return its profits to the Treasury by augmenting the number of employees on its payroll.

The employment effect that we have identified in the data is not a substitute for more traditional money supply theories, but serves as an additional determinant of the stock of high-powered money in the economy. In particular, our findings provide an explanation for what Friedman refers to as the Fed's "...extraordinary record of bureaucratic inertia, of mistakes that have not been corrected despite their widespread recognition" (p. 114). In the words of Acheson and Chant, "frequently such policy failures are judged as aberrations or are ascribed to ignorance or bad judgment" (p. 640). Our results suggest, however, that the employment effect may partially explain why the Fed has apparently been more willing to engage in expansionary than in contractionary monetary policies. As we noted earlier, contractionary open market operations reduce the Fed's profits and therefore diminish the funds available for hiring. That such a tendency exists is a consequence of having a monopoly supplier of money whose profits are confiscated by the Treasury. Competition would certainly mitigate the incentive for money producers to overinvest in amenities. However, the current monetary system may provide benefits which balance the costs associated with the use of money stock expansions to finance central bank employment. Our paper does not address such issues.

## REFERENCES

**Acheson, Keith and Chant, John F.,** "Bureaucratic Theory and the Choice of Central Bank Goals," *Journal of Money, Credit, and Banking*, May 1973, *5*, 637–55.

**Alchian, Armen A. and Kessel, Reuben A.,** "Competition, Monopoly and the Pursuit of Pecuniary Gain," in *Aspects of Labor Economics*, Universities-National Bureau of Economic Research Conference Series, No. 14, New York: Arno Press, 1962, 157–75.

**Chant, John F. and Acheson, Keith,** "The Choice of Monetary Instruments and the Theory of Bureaucracy," *Public Choice*, Spring 1972, *12*, 13–33.

______ **and** ______, "Mythology and Central

Banking, *Kyklos*, No. 2, 1973, *26*, 362–79.

**Dwyer, Gerald P., Jr.**, "Inflation and Government Deficits," *Economic Inquiry*, July 1982, *20*, 315–29.

**Friedman, Milton**, "Monetary Policy: Theory and Practice," *Journal of Money, Credit, and Banking*, February 1982, *14*, 98–118.

**Granger, C. W. J.**, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, *37*, 424–38.

**Harvey, A. C.**, *The Econometric Analysis of Time Series*, New York: John Wiley & Sons, 1981.

**Pesek, Boris P. and Saving, Thomas R.**, *The Foundations of Money and Banking*, New York: Macmillan, 1968.

**Sims, Christopher A.**, "Money, Income and Causality," *American Economic Review*, September 1972, *62*, 540–52.

**Stigler, George J.**, *The Theory of Price*, 3rd ed., New York: Macmillan, 1966.

______, "The Xistence of X-Efficiency," *American Economic Review*, March 1976, *66*, 213–6.

**Toma, Mark**, "Inflationary Bias of the Federal Reserve System: A Bureaucratic Perspective," *Journal of Monetary Economics*, forthcoming.

**Board of Governors of the Federal Reserve System**, *Annual Report*, Washington: Board of Governors, various years.

**U.S. Council of Economic Advisers**, *Economic Report of the President*, Washington: USGPO, 1982.

**U.S. Department of Commerce, Bureau of the Census**, *Historical Statistics of the United States, Colonial Times to 1970*, 2 vols., Washington: USGPO, 1975.

______, *Statistical Abstract of the United States*, Washington: USGPO, various years.

# Limited Information, Credit Rationing, and Optimal Government Lending Policy

*By* BRUCE SMITH*

A great deal of attention has been devoted to the phenomenon of credit rationing, in large part because a channel through which monetary policy can conceivably operate is the relaxation of this rationing. Surprisingly, however, the literature devoted to the microeconomic foundations of credit rationing has analyzed this phenomenon in models insufficiently rich to allow for policy analysis.[1]

This paper attempts an analysis, in a simple general equilibrium framework, of policy in the presence of credit rationing. The analysis proceeds in the context of an "informational" model of credit rationing along the lines of Dwight Jaffee and Thomas Russell (1976), but in a model sufficient to allow for money and an evaluation of government policies. The primary result which emerges is that, under fairly plausible circumstances, the government should lend as much as is demanded at some appropriately set rate of interest. At the level of abstraction of the paper, this may be interpreted as an argument that the government should operate a continuously open and unrestricted discount window. Moreover, while such an argument is reminiscent of the real bills doctrine, it is in fact stronger, since it suggests unrestricted lending to "risky" borrowers. Finally, in contrast to textbook treatments of the use of the discount window, it will be seen that in the presence of credit rationing, such a policy can be deflationary.

Under the policy suggested, the government lends enough to meet excess demand for credit at the prevailing market rate of interest. Nevertheless, it will be seen that under this policy there continues to be "rationing" in private credit markets. Moreover, the rate of interest which is charged in private credit markets is determined entirely by the rate charged by the government and the elasticity of credit supply with respect to the interest rate. This focuses discussion of the impact of further government actions on private market rates of interest and credit volume, since any such impact must occur through the alteration of this elasticity. A policy, monetary or otherwise, which does not affect this elasticity will be irrelevant insofar as the amount of credit offered is concerned.

Finally, as something of a by-product, the analysis will be seen to permit statements (on a case-by-case basis) regarding the optimal structure of credit markets. In particular, the following possibility will be demonstrated: in a credit market with asymmetric information on the part of borrowers and lenders, the monopolistic issue of liabilities, along with an open discount window, may be Pareto superior to the competitive equilibrium. In short, the analysis offers a rationale for the restriction of entry into the activity of issuing liabilities (given that a particular policy will be followed by the government). Such restriction innately gives rise to credit rationing, since it tends to restrict the supply of liabilities relative to that which would arise in a competitive credit market. However, in contrast to the suggestions of Jaffee and Russell (and others), this rationing is not desirable in and of itself, and should be eliminated through use of the discount window.

Stated succinctly, the results are as follows:

1) Under relatively unrestrictive conditions, it is welfare improving (except possibly

---

*Federal Reserve Bank of Minneapolis. The views expressed herein are solely my own and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. I thank Art Rolnick and an anonymous referee for their comments on an earlier version of this paper.

[1] Robert Barro (1976), Joel Fried and Peter Howitt (1980), Dwight Jaffee and Thomas Russell (1976), Vernon Smith (1972), Joseph Stiglitz and Andrew Weiss (1981).

for the initial old generation in the model) for the government to meet all excess demand for credit at prevailing market rates of interest.

2) This open lending policy can be deflationary.

3) Under such a policy, private market rates of interest depend only on the rate of interest charged by the government and the elasticity of credit supply with respect to interest rates.

4) Given that this policy is to be followed, monopoly issue of liabilities can be Pareto superior to competition in the issuance of liabilities.

The format of the paper is as follows. Section I discusses the Jaffee-Russell model of credit rationing as motivation for the definition of a credit rationed equilibrium. Section II presents the model employed, and considers its competitive and credit rationed equilibria. Section III discusses government lending policies and their welfare properties. It also demonstrates that under the policy suggested above, the equilibrium interest rate attained in the private credit market will depend only on supply elasticities. Finally, the possibility that an open discount window is deflationary and that monopoly issue of liabilities is desirable is demonstrated.

## I. Background

A large literature considers possible motivations for the existence of credit rationing. Here I focus on a subset of this literature which provides a motive for credit rationing based on limited, asymmetric information on the parts of borrowers and lenders. In this section I briefly discuss the model of this phenomenon produced by Jaffee and Russell. This presentation will motivate the definition employed throughout of a credit rationed equilibrium.

In the Jaffee-Russell (hereafter, J-R) model, there exist three broad groups of agents: lenders; safe borrowers; and risky borrowers. Safe borrowers never default on loans, whereas risky borrowers default with positive probability. To simplify matters, it will be assumed that this probability increases with the volume of lending. Lenders are assumed to be unable to distinguish *ex ante* between safe and risky borrowers unless their economic behavior differs. Lenders are also expected profit maximizers who incur no transactions costs. Moreover, there is free entry into lending. Hence the condition defining credit supply is a zero profit condition.

It is an assumption of the J-R analysis that risky borrowers will wish to mimic safe borrowers to obtain more favorable rates of interest.[2] Under this assumption there is only a single credit market, and the supply curve of credit in the market is upward sloping. This is because as loan sizes increase, the probability of default *ex ante* increases for each loan (since safe and risky borrowers are indistinguishable *ex ante*), and hence interest rates must rise to make expected profits zero.

The situation in the credit market is depicted in Figure 1. The projection of the Lagrangian for safe borrowers into this space is the indifference curve *II*, which attains a maximum as a function of loan size where it intersects the loan demand schedule. Point *C* is the competitive equilibrium position in this market.

The motivation for credit rationing in this setting is to note that, given the implied departure from an Arrow-Debreu setting, the exercise of monopoly and/or monopsony power can be Pareto improving.[3] To see this, consider point *E* in Figure 1. At point *E*, safe borrowers are as well off as at the competitive equilibrium and lenders make positive expected profit. In fact, any point in the region *CED* represents a joint improvement for safe borrowers and lenders. Under restrictions on the preferences of risky borrowers, their welfare will also be improved by movement away from the competitive equilibrium.

With free entry into lending, it is true that a point like *E* cannot be an equilibrium. The J-R argument is then, that at any point other than *M*, where an indifference curve is tangent to the supply curve, it is possible for some lenders to offer borrowers an interest rate-loan package which they prefer to the package offered by other lenders, and which

---

[2] In fact, this represents a restriction on preferences.

[3] See, for example, Peter Diamond (1980).

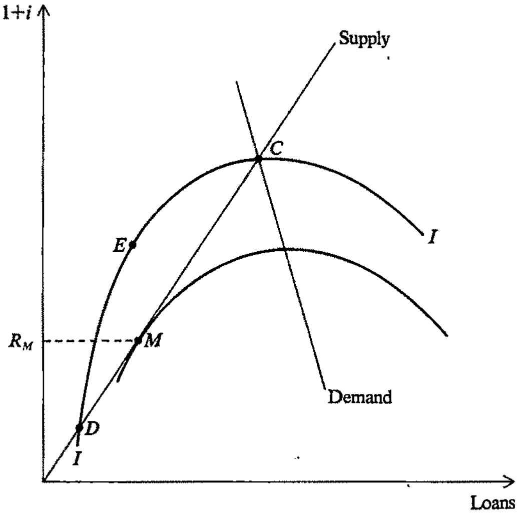


FIGURE 1

makes positive expected profit. Thus, only $M$ leaves no incentive for some lenders to offer a different interest rate-loan package, and hence $M$ is the equilibrium of a particular game.

To motivate the definition employed below of a credit rationed equilibrium, I note the following. First, at the interest rate $R_M$, notional demand exceeds credit supply. Hence there is credit rationing in this equilibrium. Second, the claim made is that a credit rationed equilibrium occurs where safe borrowers' indifference curves are tangent to the offer curves of lenders. This is simply the monopsonistic equilibrium for this market, which motivates the definition of a credit rationed equilibrium employed below.[4]

Prior to concluding this section, more should be said about this notion of a credit rationed equilibrium. In particular, it could be argued with some validity that the coincidence of a credit rationed equilibrium with the monopsonistic equilibrium is something of an artifact of the J-R model. However, this is not to say that the monopsonistic equilibrium is not of interest. It is, in fact, of interest in its own right, since as will be seen, under appropriate government policies this equilibrium can be Pareto superior to the competitive equilibrium in a fairly general credit market. Moreover, the monopsonistic equilibrium features borrowers who would like to obtain more credit than they can at the market rate of interest; that is, it generates credit rationing as an outcome. Thus, either as an extension of the J-R analysis, or as an equilibrium concept of interest in its own right, I focus below on monopsonistic equilibria as a source of credit rationing.

## II. The Model and Its Equilibria

### A. *Specification*

Let us consider a model sufficiently rich to permit a welfare analysis of government policies in the presence of credit rationing. To this end, I present a model in the spirit of Section I, but which is enhanced in several respects. The economy considered exists in a discrete time world with time indexed by $t = 0, 1, \ldots$. This economy is populated by a sequence of two-period lived, overlapping generations, and an initial old generation at $t = 0$. There is a single nonstorable consumption good at each date, as well as a fixed (for all time) stock of money, $M$. Consumption loans are also permitted between agents. Throughout, attention will be confined to equilibria in which fiat money has value.

With the possible exception of the initial old, each generation consists of three types of agents, indexed by $i = 1,2,3$. Given our assumptions on endowment patterns, type 1 agents lend (and hold money), type 2 agents borrow and never default, and type 3 agents borrow and default with positive probability. Type $i$ agents have preferences described by the twice-continuously differentiable, strictly concave utility functions $U_i(C_1, C_2)$, with $U_i(\ )$ strictly increasing in each argument at every point in the consumption set, $\mathfrak{R}^2_+$. The arguments of $U_i$ are consumption when young and old, respectively. Notation will not be required for the consumption of different agents, for the dating of consumption, or for consumption in alternate states of nature.

---

[4] Note that the J-R argument is that banks do not lend as much as borrowers demand at prevailing interest rates so as to replicate the monopsonistic equilibrium outcome in credit markets. While it is difficult to accept this as an explanation of real world credit rationing, this is a commonly accepted rationale for its existence. Hence we proceed along these lines.

Endowment streams are as follows: type 1 agents have endowment $b$ of the consumption good when young, and zero when old; type $i$ agents ($i = 2, 3$) have zero endowments when young. Type 2 agents have endowment $w_2$ when old (with certainty). Type 3 agents have endowment $w_3 > 0$ when old with probability $p$, and endowment zero (so that they default on loans) with probability $1 - p$. If agents receive positive endowments when old, they do not default on loans. Agents have no endowments of either money or consumption loans, except that the initial old are endowed at $t = 0$ with the entire money stock.

Consumption loans are made and repaid in the consumption good, which is chosen as the numeraire. Money trades for the consumption good at rate $S_t$ at time $t$. Throughout I restrict consideration to steady-state equilibria and hence often omit subscripts on $S$. If lenders can distinguish between safe and risky borrowers, then I denote loans to safe borrowers by $x_1$, and to risky borrowers by $x_2$. Safe borrowers repay amount $R_1x_1$ when old if they borrow $x_1$, and risky borrowers repay $R_2x_2$ if they do not default. If lenders cannot make this distinction *ex ante*, then I denote loans to both types of borrowers by $x$, and agents repay amount $Rx$ if they do not default. Finally, for notational convenience, it is assumed that each generation (except possibly the initial old) consists of a single agent of each type.

### B. *Competitive Equilibrium*

We are now ready to define two competitive equilibria for this economy (two due to the nature of the informational asymmetry in the model). Let us begin with a discussion of this asymmetry, and motivate in this way the two types of competitive equilibria.

Note first that when young, type 2 and 3 agents have identical endowments, and in terms of currently realized states of nature appear identical prior to making any economic decisions. Hence, assume that type 1 agents cannot distinguish between young type 2 and type 3 agents unless their economic decisions differ. (Each agent knows his own future endowment as a function of future states of nature.) Since agents are distinguishable when young only through their behavior, agents with $i = 3$ may choose either to reveal or not to reveal their type. If they choose not to reveal, they must mimic the behavior of type 2 agents. Hence there are two possibilities as regards competitive equilibria; these may be either revealing or nonrevealing as regards each agent's default probability.

We are now ready to define a full-information (revealing) and a limited-information (nonrevealing) competitive equilibrium. As attention is confined to steady states, definitions need not be stated in terms of sequences of values. The definitions here differ from standard definitions of competitive equilibrium, and from the Jaffee-Russell definition, because in equilibrium, type 3 agents either reveal their indices through their actions or not. If these agents choose to reveal themselves, or not to reveal themselves, we require this choice to be incentive compatible with actual outcomes. This adds an additional condition to the definitions of equilibrium.

DEFINITION: *A full-information (steady-state) competitive equilibrium (FIE) is a set of consumption pairs* $(C_1, C_2)$ *for agents* $i = 1, 2, 3$, *a pair* $(R_1, R_2)$, *a pair* $(x_1, x_2)$ *with* $x_1 \neq x_2$, *and a value* $S \geqslant 0$ *such that*

(*i*) *for type* 1 *agents* $(C_1, C_2, x_1, x_2, M)$ *solves* $\max EU_1(C_1, C_2)$ *subject to* $C_1 \leqslant b - x_1 - x_2 - SM$; $C_2 \leqslant R_1x_1 + R_2x_2 + SM$ *in nondefault states, and* $C_2 \leqslant R_1x_1 + SM$ *in default states, where* $E$ *is the expectations operator, with expectations taken with respect to the objective probability of default.*

(*ii*) *For type* 2 *agents* $(C_1, C_2, x_1)$ *solves* $\max U_2(C_1, C_2)$ *subject to* $C_1 \leqslant x_1$; $C_2 \leqslant w_2 - R_1x_1$;

(*iii*) *for agents with* $i = 3$ $(C_1, C_2, x_2)$ *solves* $\max EU_3(C_1, C_2)$ *subject to* $C_1 \leqslant x_2$; $C_2 \leqslant w_3 - R_2x_2$ *in nondefault states and* $C_2 = 0$ *in default states.*

(*iv*) *The values $R_1$, $R_2$ and S imply that loan and money markets clear.*

(*v*) *Incentive compatibility*:

$$pU_3(x_2, w_3 - R_2x_2) + (1-p)U_3(x_2, 0)$$
$$\geqslant pU_3(x_1, w_3 - R_1x_1) + (1-p)U_3(x_1, 0).$$

This is simply the (parameter contingent) competitive equilibrium of Edward Prescott and Robert Townsend (1981), or of my earlier paper.

DEFINITION: *A limited information (steady-state) competitive equilibrium, or LIE, is a set of pairs $(C_1, C_2)$ for each type of agent, a pair $(R, x)$, and a value $S \geqslant 0$ such that*

(*vi*) *for agents with $i = 1$, $(C_1, C_2, x, M)$ solves $\max EU_1(C_1, C_2)$ subject to $C_1 \leqslant b - 2x - SM$, $C_2 \leqslant 2Rx + SM$ in nondefault states, and $C_2 \leqslant Rx + SM$ in default states*;

(*vii*) *for agents with $i = 2$, $(C_1, C_2, x)$ solves $\max U_2(C_1, C_2)$ subject to $C_1 \leqslant x$, $C_2 \leqslant w_2 - Rx$,*

(*viii*) $$pU_3(x, w_3 - Rx) + (1-p)U_3(x, 0)$$
$$\geqslant pU_3(x_2, w_3 - R_2x_2) + (1-p)U_3(x_2, 0),$$

(*iv′*) *the values of R and S imply loan and money market clearing.*

Condition (*viii*) requires that if type 3 agents choose not to reveal their indices, then they must prefer the *LIE* outcome to the outcome which obtains if they do reveal their indices. If there are multiple values of $R_2$ and $x_2$ satisfying (*i*)–(*iii*) and (*iv*), (*viii*) should be taken to hold for any of these values. This definition, then, is simply the J-R competitive equilibrium, with an incentive compatibility condition added.

Before proceeding, it is useful to clarify the relationship between these definitions of equilibrium and the standard definitions of equilibrium employed (for instance) in the insurance literature. In keeping with the Rothschild-Stiglitz (1976) notion of equilibrium, we might think of some agents offering loan-interest rate pairs.[5] If it is possible for lenders to distinguish safe from risky borrowers, they offer distinct interest rate-loan combinations. If not, these agents offer a single such pair. However, in contrast to the Rothschild-Stiglitz definition of equilibrium, we require any interest rate-loan pair offered to be market clearing at the stated interest rate. Relative to the Rothschild-Stiglitz definition (where any price-quantity pair can be offered), this makes it more difficult for a revealing equilibrium to exist, and less difficult for a nonrevealing equilibrium to exist. As we wish to consider nonrevealing equilibria (in keeping with the J-R analysis), some such modification of the equilibrium concept is required, with the definition above being perhaps the simplest one.

### C. *Existence of Equilibrium*

In keeping with the J-R model of credit rationing, attention will be confined to *LIE*s. Hence it is required that equilibria satisfy, among other conditions, condition (*viii*). This requirement implies that a *LIE* (and, in general, any competitive equilibrium) need not exist under normal regularity assumptions.[6] Therefore, I establish that the concept of a *LIE* is nonvacuous by means of an example. (Moreover, the example also establishes that there exist economies with a *LIE*, and with no competing *FIE*s.)

*Example 1:* Let preferences be given by

$$U_1 = C_1 + dC_2 - (\phi/2)(C_2)^2$$

$$U_2 = \ln C_1 + \ln(1 + C_2)$$

$$U_3 = \ln C_1 + \ln(1 + C_2),$$

with $d = 2$, $w_2 = 7$, $w_3 = 5$, $p = 1/2$, and $\phi = 1/10$. As long as it is chosen to guarantee feasibility, the value of $b$ is immaterial.

---

[5] For a more formal definition of *FIE*s and *LIE*s along these lines, see my earlier paper.

[6] See my earlier paper.

Let us first establish that this economy has no *FIE*, steady state or otherwise. (In fact, revelation of type is impossible at any date.) Given the set of preferences above, condition $(v)$ is

$$\ln\left(\frac{1+w_2}{2R_1}\right)+p\ln\left[1+w_3-\left(\frac{1+w_2}{2}\right)\right]$$
$$\leqslant \ln\left[\frac{1+w_3}{(1+p)R_2}\right]+p\ln\left[1+w_3-\left(\frac{1+w_3}{1+p}\right)\right].$$

For our parameter values, this condition reduces to $R_2 \leqslant R_1$. Since type 1 agents are risk averse, this is impossible in equilibrium. Hence this economy has no *FIE*. However, it is readily verified that a *LIE* does exist for this economy. This has $R=10/7$, $x=2.8$, and $S_t=4/M$. Note that the levels of expected utility attained in this equilibrium are $EU_1=b+5.2$, $EU_2=2.416$, and $EU_3=1.376$.

It is the case, then, that there exist economies with no *FIE*s, and for which the concept of a *LIE* is nonvacuous. I now proceed in defining a credit rationed equilibrium.

## D. *Credit Rationed Equilibria*

The definition of a credit rationed equilibrium is motivated by recalling that the J-R credit rationed equilibrium is the limited-information monopsonistic equilibrium for their economy. Only a limited-information equilibrium is defined here for reasons discussed below.

Denote the optimal choice of $x$ for type 1 agents at any given set of prices by $x_s^*((S_{t+1})/S_t, R, b) = x_s(R, b)$, since our focus on steady states implies $S_{t+1}/S_t=1$. Then

DEFINITION: *A limited-information (steady-state) monopsonistic equilibrium (LIME), or credit rationed equilibrium, satisfies conditions* $(vi)$, $(viii)$,

$(ix)$ *the 4-tuple* $(C_1, C_2, x, R)$ *for type* 2 *agents solves* $\max U_2(C_1, C_2)$ *subject to* $C_1 \leqslant x$; $C_2 \leqslant w_2 - Rx$; $x = x_s(R, b)$,

$(x)$ *the value S clears the money market.*

The reason for defining only a limited information version of this equilibrium is as follows. The Jaffee-Russell analysis is such that only the monopsony power of type 2 agents is relevant. I maintain a version of this by assuming that only type 2 agents can exercise monopsony power.[7] Given this assumption, there is no full-information monopsonistic equilibrium, (or more precisely, any such equilibrium is a *FIE*). This is because, with full information, money holdings are a perfect substitute for loans to type 2 agents in the portfolios of type 1 agents. Hence safe borrowers face a perfectly elastic supply curve of credit and have no monopsony power. Therefore, the relevant interest rate in the incentive compatibility constraint is the full-information rate of interest $R_2$, which appears in $(viii)$.

## E. *Existence of a LIME*

In general, a *LIME* need not exist for either of two reasons. One is the conventional reason for nonexistence of monopolistic/monopsonistic equilibria: reaction correspondences need not be convex valued. The second is that just as a competitive equilibrium need not exist as a result of imposing the incentive compatibility conditions, this may result in nonexistence of a monopsonistic equilibrium as well. Given that a *LIME* need not exist, the remainder of this section is devoted to demonstrating that the notion of a *LIME* is nonvacuous. This is done by means of an example.

*Example 2:* Preferences and parameter values are as for example 1. For convenience, note that

$$x_s(R,b)=\frac{(1+p)R-2}{(1-p)p\phi R^2}$$

for this example, and that

$$D_R x_s(R,b)=(1-p)p\phi R[4-(1+p)R] \times\left[(1-p)p\phi R^2\right]^{-2},$$

---

[7] Even in the J-R analysis this is an implicit assumption. Otherwise there would be no need for limited information as a rationale for credit rationing. Moreover, the full-information monopsonistic equilibrium is of no particular interest. This is discussed further below.

where $D_R$ denotes differentiation with respect to $R$. The supply curve of credit is upward sloping so long as $R < 4/(1+p)$, which is the relevant range for this example. It is also the case that over the range where the supply curve is upward sloping, it is a strictly concave function of $R$, so that second-order conditions for the problem (*ix*) are satisfied for this example.

It is tedious but straightforward to verify that the values of $R$ and $x$ satisfying (*ix*) are $R = 1.425 < (10/7)$, and $x = 2.709$. Since $10/7$ is approximately 1.429, for this example credit rationing results in about a 1 percent reduction in the interest rate $(R-1)$. Finally, we must verify that the incentive compatibility condition (*viii*) is satisfied. Note that if $pR_2 = 1$, then loans to type 3 agents under full information have equal expected return to the holding of money, or safe loans. Hence, since type 1 agents are risk averse, $(1/p)$ represents a lower bound for $R_2$. Then

$$
\begin{aligned}
& pU_3(x_2, w_3 - R_2x_2) + (1-p)U_3(x_2, 0) \\
& \leqslant pU_3\left(x_2, w_3 - \frac{x_2}{p}\right) + (1-p)U_3(x_2, 0) \\
& = ln4 - ln2 + (1/2)\,ln2 = 1.04 \\
& < 1.377 = ln(2.709) + (1/2)\,ln(6 - 3.86) \\
& = ln\,x + p\,ln(1 + w_3 - Rx).
\end{aligned}
$$

This verifies that type 3 agents do prefer not revealing themselves to revelation, so the *LIME* is incentive compatible, and $S_t = (4.21/M)\forall t \geqslant 0$, $R = 1.425$ is, in fact, a *LIME*.

Note that the levels of expected utility attained under the *LIME* are $EU_1 = b + 5.186$, $EU_2 = 2.418$, and $EU_3 = 1.377$. Thus, the *LIE* and the *LIME* are not Pareto comparable.

Having established that the concept of a *LIME* is nonvacuous, we may now consider government policies for economies where a *LIME* exists.

## III. Policy

In this section, it is demonstrated that under plausible circumstances, it is welfare improving for all agents (except possibly the initial old) if the government meets all excess demand for credit at prevailing market interest rates. This is true despite the fact that it involves operating an open "discount window" for risky borrowers. In addition, it is demonstrated that this policy has the effect of making free-market interest rates (and credit volume) invariant to anything other than the elasticity of credit supply with respect to the rate of interest. Finally, it will be seen that this policy may be deflationary, and that in its presence, the monopoly issue of liabilities may be Pareto superior to their competitive issue.

The scheme of the section is as follows. Part A establishes sufficient conditions for the desirability of this policy under the assumption that a *LIME* exists before and after its implementation. Part B establishes the invariance argument for interest rates under the same assumption. Part C establishes that these results are nonvacuous, that is, that an equilibrium can exist under this policy. It also presents the possibility of deflation and of desirable monopoly-liability issue under the policy.

### A. *Government Lending*

Here it is established that the following policy improves the welfare of all (but possibly the initial old) agents under fairly general circumstances: (a) the government lends as much as agents desire to borrow at the *LIME* interest rate $R^*$. (b) This policy is financed by appropriate lump sum taxes-cum-transfers $T_i(s)$ on agent $i$ in state $s$ ($s = 0$ is the nondefault state, $s = 1$ the default state) which balance the government budget in each state and period.

This intervention by the government will alter the equilibrium which obtains. Therefore an additional definition is required.

DEFINITION: *A post-intervention LIME is an array of pairs* $(C_1, C_2)$, *a pair* $(\hat{R}, \hat{x})$, *and a value S satisfying*

(*xi*) *for type* 1 *agents the* 4-*tuple* $(C_1, C_2, x, SM)$ *solves* $\max EU_1(C_1, C_2)$ *subject to* $C_1 \leqslant b + T_1 - 2x - SM$; $C_2 \leqslant 2Rx +$

*$SM$; $s=0$, $C_2 \leqslant Rx+SM$; $s=1$, $T_1$ parametric.*

Denote the solution for $x$ by

$$x = x_s(R, b+T_1).$$

*(xii) For type 2 agents the 4-tuple $(C_1, C_2, \hat{x}, \hat{R})$ solves $\max EU_2(C_1, C_2)$ subject to $C_1 \leqslant \hat{x}$; $C_2(s) \leqslant w_2 - R^*\hat{x} + (R^* - \hat{R})x_s + T_2(s)$; $s = 0, 1$; $x_s = x_s(\hat{R}, b+T_1)$; $T_2(s)$ parametric, $s=0,1$.*

*(xiii) $pU_3[\hat{x}, w_3 - R^*\hat{x} + (R^* - \hat{R})x_s + T_3] + (1-p)U_3(\hat{x}, 0) \geqslant pU_3(x_2, w_3 - R_2x_2 + T_3) + (1-p)U_3(x_2, 0)$, where $R_2$ and $x_2$ are to be understood as values satisfying (i)–(iii) and (v), given the values $T_i(s)$.*

*(xiv) The value $S$ clears the money market.*

In short, type 2 agents borrow as much as desired in private credit markets, exploiting their monopsony power in the process. They then borrow the difference between this quantity and their notional demand from the government at the interest rate $R^*$. Type 3 agents mimic the actions of type 2 agents in each case.

The primary result is as follows:

PROPOSITION 1: *If* (a) *$\hat{x} \geqslant x_s(R^*, b)$ and* (b) *$x_s(R^*, b)$ is less than the notional demand of type 3 agents at the interest rate $R^*$, then except for the initial old, it is Pareto improving for the government to lend as much as demanded at the pre-intervention LIME interest rate.*[8]

To establish the proposition, and to derive the sufficiency of conditions (a) and (b) simultaneously, I begin by defining $V_i(S_{t+1}/S_t, R, y_1, y_2)$ to be the indirect utility function of type $i$ agents who are young at $t$, where $y_j$ is income in the $j$th period of life of agent $i$. The focus on steady states implies $S_{t+1}/S_t = 1$, and thus this argument is omitted in the notation. In addition, as I construct the proof of the proposition, each agent will have no endowment or income in one period of life. This argument is suppressed as well.

PROOF of Proposition 1:

Let $R^*$ denote the *LIME* interest rate prevailing at the date of the initial policy implementation, which we take to be $t=0$. The term $\hat{R}$ denotes the *LIME* interest rate in the presence of this policy, which consists of the government making all loans demanded at $R^*$, with the level of transfers implied taken as parametric by agents. The method of proof is to construct transfers in a way which satisfies the proposition. To do this, consider agents with $i=1$. Since in both the pre- and post-intervention *LIME* type 1 agents are on their supply curves, the change in their utility from the program is $V_1(R^*, b) - V_1(\hat{R}, b+T_1)$, where transfers to type 1 agents are $T_1$, which are made when young and are *not* state dependent.[9] Taking a first-order approximation and equating this difference to zero, we have

$$D_RV_1(R^*, b)(R^* - \hat{R}) - D_yV_1(R^*, b)T_1 = 0.$$

For "small" levels of credit rationing this is a reasonable approximation. Solving for $T_1$, we obtain

$$T_1 = \frac{D_RV_1(R^*, b)(R^* - \hat{R})}{D_yV_1(R^*, b)}.$$

---

[8] The statement holds for any pre-intervention interest rate which results in credit rationing or, in other words, holds for a wide variety of equilibrium concepts other than that of a *LIME*.

[9] We require in all cases transfers to be received only in the period (and states) in which the agent in question has a positive endowment. Otherwise we are allowing government redistribution of income which is not feasible for agents, and the proposition is trivial. In addition, it is this requirement which makes government transfers informationally feasible. It is an (implicit) assumption that the receipt of each agent's second period endowment is public information. Thus, while type-dependent transfers to type 2 and 3 agents are not possible when these agents are young (they are indistinguishable), they are possible when these agents are old and the second-period endowment has been realized (so long as $w_2 \neq w_3$).

Noting that the intertemporal version of Roy's Identity for this problem implies

$$\frac{D_R V_1(R^*,b)}{D_y V_1(R^*,b)}=\frac{2x_s(R^*,b)}{R^*},$$

we have $T_1=\dfrac{2x_s(R^*,b)(R^*-\hat{R})}{R^*}$.

This transfer guarantees that all interest rate changes resulting from the program are income compensated from the point of view of type 1 agents.

Consider next agents with $i=3$. Under the program the change in their utility is

$$pU_3\big[\hat{x},w_3-\hat{R}\hat{x}+(R^*-\hat{R})x_s(\hat{R})+T_3(0)\big]$$
$$+(1-p)U_3(\hat{x},0)$$
$$-pU_3(x^*,w_3-R^*x^*)-(1-p)U_3(x^*,0),$$

where $\hat{x}$ is the post-intervention *LIME* value of $x$, $x^*\equiv x_s(R^*,b)$, $T_3(0)$ are transfers to type 3 agents in state $s=0$, and $T_3(1)\equiv 0$. For now, we simply state that $T_3(0)=-(R^*-\hat{R})x_s(\hat{R},b+T_1)$, and return to the consideration of type 3 agents below.

In order for the government budget to balance, all remaining transfers must be made to type 2 agents. These will be state dependent transfers which type 2 agents receive when old. These are denoted $T_2(s)$, and by the budget balance condition, are defined by

$$(1)\quad T_2(0)=2(R^*-1)\big[\hat{x}-x_s(\hat{R},b+T_1)\big]-T_1-T_3(0);$$

$$(2)\quad T_2(1)=(R^*-2)\big[\hat{x}-x_s(\hat{R},b+T_1)\big]-T_1.$$

Equation (1) states that in nondefault states, the government lends the difference between what agents wish to borrow at the interest rate $R^*$, and the amount offered in the marketplace at $\hat{R}$, the post-intervention *LIME* interest rate. In addition, two loans of the same amount are repaid. Finally, the government pays out $T_1$ to compensate type 1 agents, and collects $T_3$ from type 3 agents. The net profit to the government then accrues to type 2 agents (who are distinguishable *ex post* when second-period endowments have been realized). Equation (2) differs in that only one loan is repaid when $s=1$ and $T_3(1)=0$.

I now demonstrate conditions sufficient for the utility of type 2 agents to increase. First note that, since the problem (*xii*) is separable, the maximized value of (*xii*) is (with an obvious notation)

$$V_2\big[R^*,w_2+T_2(0)+(R^*-\hat{R})x_s,w_2+T_2(1)+(R^*-\hat{R})x_s\big],$$

where $x_s\equiv x_s(\hat{R},b+T_1)$. If this is greater than or equal to $V_2(R^*,w_2)$, the utility of type 2 agents does not worsen as a result of the program. Defining $\tilde{T}_2(s)=T_2(s)+(R^*-\hat{R})x_s(\hat{R},b+T_1)$ and taking a first-order approximation about $\tilde{T}_2(s)=0$; $s=0,1$, gives

$$pD_yV_2(R^*,w_2)\tilde{T}_2(0)+(1-p)D_yV_2(R^*,w_2)\tilde{T}_2(1)\geqslant 0,$$

or

$$D_yV_2(R^*,w_2)E\tilde{T}_2(s)\geqslant 0$$

as the condition under which type 2 utility increases. This is equivalent to

$$(3)\quad E\tilde{T}_2(s)=\big[(1+p)R^*-2\big](\hat{x}-x_s)+(R^*-\hat{R})\left[(1+p)x_s-\frac{2x^*}{R^*}\right]\geqslant 0.$$

Now note that if $(1+p)R^*/2=1$, the expected return on loans is unity. Since lenders are risk averse, $2/(1+p)$ constitutes a lower bound on $R^*$. Therefore, since $\hat{x}>x_s(\hat{R},b+T_1)$,[10] the first term on the right-hand side of equation (3) is positive. It is also the case that $R^* > \hat{R}$. Then if $x_s \geqslant x^*$, it is immediate that $E\tilde{T}_2(s) \geqslant 0$, and the utility of type 2 agents is increased by the program. However, it will often be the case that $x_s(\hat{R}, b+T_1) < x_s(R^*, b)$, so that the sign of (3) is ambiguous. We now investigate conditions under which (3) holds.

---

[10] This follows from the fact that the economy was initially in a *LIME* and from type 2 agents' maximization problem in the presence of government policy.

To begin, note that $\hat{R} > 2/(1+p)$. Therefore, if the second term on the right-hand side of (3) is negative,[11]

$$E\tilde{T}_2(s) > [(1+p)R^* - 2](\hat{x} - x_s)$$
$$-\left(R^* - \frac{2}{1+p}\right)\left[\frac{2x_s(R^*,b)}{R^*} - (1+p)x_s\right].$$

Then $E\tilde{T}_2(s) > 0$ if

$$\left(R^* - \frac{2}{1+p}\right)\Bigg\{(1+p)(\hat{x} - x_s)$$
$$-\frac{2x_s(R^*,b)}{R^*} + (1+p)x_s\Bigg\} = \left(R^* - \frac{2}{1+p}\right)$$
$$\left[(1+p)\hat{x} - \frac{2x_s(R^*,b)}{R^*}\right] \geqslant 0.$$

Since $R^* > 2/(1+p)$, this is equivalent to

$$(1+p)\hat{x} - \frac{2x_s(R^*,b)}{R^*} \geqslant 0.$$

Now note that since $R^* > 2/(1+p)$,

$$(1+p)\hat{x} - 2x_s(R^*,b)/R^*$$
$$> (1+p)\hat{x} - (1+p)x_s(R^*,b).$$

Thus condition (a) in Proposition 1 ($\hat{x} \geqslant x_s(R^*, b)$) is sufficient for the utility of type 2 agents to increase as a result of the program. This condition is easily interpreted as type 2 utility increases if the changes in income for type 2 agents implied by the program would not result in notional demand being less than the original level of credit supplied in the *LIME*. In other words, $\tilde{T}_2(s)$ may not convert the situation of excess notional demand into excess supply at $R^*$.

Consider now the welfare of type 3 agents. Their welfare increases if

$$(4)\quad EU_3[\hat{x}, w_3(s) - R^*\hat{x}]$$
$$> EU_3[x^*, w_3(s) - R^* x_s(R^*,b)],$$

where $x^*$ is the pre-intervention *LIME* value. To a first-order approximation, equation (4) holds if

$$(5)\quad [\hat{x} - x^*]E(D_{C_1}U_3)$$
$$> R^*[\hat{x} - x^*]E(D_{C_2}U_3).$$

This is equivalent to condition (b) of Proposition 1:

$$E(D_{C_1}U_3)/E(D_{C_2}U_3) \gtrless R^* \text{ as } \hat{x} \gtrless x^*.$$

If $E(D_{C_1}U_3)/E(D_{C_2}U_3) > R^*$, then the notional demand of type 3 agents at $R^*$ exceeds $x^*$. If (a) and (b) are satisfied, then the policy of Proposition 1 increases the welfare of both type 2 and 3 agents. Condition (b) has the plausible interpretation that if (a) is satisfied, risky lenders would borrow more than they actually do if their activity was not constrained by the actions of type 2 agents. In other words, if condition (a) is satisfied, (b) implies that type 3 agents (as do type 2 agents) face credit rationing in a *LIME*. This is a plausible condition.

It remains to consider the financing of the program at $t=0$. At all dates $t>0$, the program is either self-financing, or financed by lump sum taxes on old agents with $i=2$. It is assumed that at $t=0$, the program is financed by a lump sum tax on the initial old generation. Sufficient conditions for their utility not to fall as a result amount to the price of money rising sufficiently to offset this tax. This, in turn, requires sufficiently large increases in money demand as $\hat{R}$ falls below $R^*$. Such a condition will not typically be satisfied.[12] Thus, the initial old must generally be exempted from the proposition.

---

[11] Where strict inequality holds because $\hat{R} > 2/(1+p)$. This is easily seen from the conditions in Section III, Part B, for the optimal value of $\hat{R}$.

[12] Although it will be shown below that it can be.

The notions behind Proposition 1 can be simply stated as follows. With reference to Figure 1, if safe borrowers are allowed to borrow according to their notional demand at $R_M$, they will clearly be made better off. If risky borrowers are also credit rationed at $R_M$, this represents a movement in a utility-improving direction for them if it is not too large. Moreover, since incremental government lending reduces the use of private credit markets, the variability in second-period consumption tends to be reduced for lenders.[13] Counteracting this last tendency, however, is the fact that the interest rate received by private lenders is reduced by the policy. Proposition 1 establishes conditions which are sufficient to guarantee that, even if lenders are made worse off by an open discount window, borrowers gain sufficiently from the policy to compensate them. Moreover, it will be noted that the policy change is welfare improving under these conditions even though the government operates under the same informational limitations as private lenders.

The proposition has been established, then, under several assumptions. The first is that credit rationing is small. This assumption is not strictly required for the proposition, as I have established much stronger conditions than required in the presence of small credit rationing. The second is that a post-intervention *LIME* exists. This is a nontrivial assumption even though, by hypothesis, a pre-intervention *LIME* exists. The problem essentially rests with the fact that, taking transfers as parametric, the corresponding full-information interest rate $R_2$ may change so as to increase the expected utility associated with revelation for type 3 agents by more than utility increases for these agents in a *LIME*. In other words, under the transfer scheme implied, a post-intervention *LIME* need not be incentive compatible even though the pre-intervention *LIME* was. Thus, it remains to demonstrate that there exist economies for which the policy of Proposition 1 is feasible. This is demonstrated in Part C below.

### B. *The Post-Intervention LIME*

Here I demonstrate that under the policy of Proposition 1, private market rates of interest and credit volume are invariant with respect to a large class of potential government actions. I also indicate what these market quantities do depend on. The result obtained is stated as

PROPOSITION 2: *Under the policy of Proposition 1, $\hat{R}$ depends only on $R^*$ and the elasticity of credit supply with respect to $R$.*

PROOF:

Consider the first-order condition for $\hat{R}$ associated with the problem (*xii*). This is

$$pD_{C_2}U_2|_{s=0}\left[x_s-(R^*-\hat{R})D_Rx_s\right]+(1-p)$$
$$\times D_{C_2}U_2|_{s=1}\left[x_s-(R^*-\hat{R})D_Rx_s\right]=0.$$

In order for this condition to be satisfied, we require

$$x_s-(R^*-\hat{R})D_Rx_s(\hat{R},b+T_1)=0$$

or

$$\frac{\hat{R}D_Rx_s(\hat{R},b+T_1)}{x_s}\equiv\eta_s(\hat{R})=\frac{\hat{R}}{R^*-\hat{R}}, \tag{6}$$

where $\eta_s$ is the elasticity of credit supply with respect to $R$. In short, (6) states that type 2 agents simply minimize expenditures by choice of $\hat{R}$, and establishes the proposition.

Proposition 2 demonstrates that while a wide class of policies *will* affect private credit markets, a large class of policies will have no such effect. As an example, if the analysis were extended to include safe bonds, open market operations would have no effect on private credit markets. Such irrelevance is, of course, not unique to this setting. However, the main point of Proposition 2 is that it demonstrates the exact focal point of any discussion of the impact of further policy actions on credit markets under government policies which are not clearly suboptimal. For such actions to be effective, it must be demonstrated that there is some impact of these policies on supply elasticities of credit.

---

[13] This is irrelevant in the Jaffee-Russell analysis because lenders are risk neutral. It also does not appear in the demonstration of Proposition 1 because of the focus on small credit rationing.

### C. *An Example*

Here I demonstrate that Propositions 1 and 2 are nonvacuous, or more specifically, provide a sample economy for which they hold (i.e., which has a post-intervention *LIME*). As a by-product of this demonstration, it will be seen that the policy of Proposition 1 is deflationary for this economy. It is also Pareto improving over both the *LIME*, *and the LIE* for the economy. This establishes the possibility that, given a commitment to run an open discount window at an appropriately set rate of interest, the monopoly issue of liabilities Pareto dominates their competitive issue. In other words, it is possible that restrictions on entry into the activity of issuing liabilities are desirable. This should not be taken as a generalizable result, however.

*Example 3.* The economy is as for examples 1 and 2. As a restatement:

$$T_1 = x_s(R^*, b)(R^* - \hat{R})$$

$$T_3(0) = -(R^* - \hat{R})x_s, T_3(1) = 0$$

$$T_2(0) = 2(R^* - 1)(\hat{x} - x_s) - T_1 - T_3(0)$$

$$T_2(1) = (R^* - 2)(\hat{x} - x_s) - T_1.$$

As is clear from the preferences of type 1 agents, the supply schedule of credit is unaffected by the presence of transfers. This supply curve, repeated for convenience, is

$$x_s(R) = \frac{(3/2)R - 2}{(1/40)R^2} = \frac{60R - 80}{R^2}$$

for the parameters of our example. The post-intervention *LIME* value of $R$, $\hat{R}$, maximizes

$$(R^* - \hat{R})x_s(\hat{R}) = (1.425 - \hat{R})\left(\frac{60\hat{R} - 80}{\hat{R}^2}\right).$$

The optimal $\hat{R}$ and associated level of credit supply are $\hat{R} = 1.378$, $1.41 = x_s(\hat{R})$. It is tedious but straightforward to derive other equilibrium values. These are $\hat{x} = 2.86$; $T_2(0) = 1.34$; and $T_2(1) = -1.01$. It is also possible to verify that the change in the steady-state value of real balances between this equilibrium and the equilibrium of example 2 is $\Delta SM = 2.88$.

It remains to verify that the post-intervention *LIME* is incentive compatible. To see that it is, compute $EU_3 = 1.41$. It was established in example 2 that an upper bound for $EU_3$ under revelation was 1.04, so the post-intervention *LIME* is incentive compatible. This completes the demonstration of nonvacuousness.

Also note that the example establishes that this "liberal use of the discount window" can be deflationary (real balances rise with the stock of money constant). Hence it is *not* the case in the presence of credit rationing that liberal use of the discount window generally constitutes an inflationary policy action.[14]

Finally, I use the example to demonstrate that this policy can be strictly welfare improving. Note first that $\Delta SM = 2.88 > \hat{x} = 2.86 > \hat{x} - x_s(\hat{R})$, so that the real balances of the initial old rise by more than enough to compensate them for the tax imposed to finance initial government lending.

Second, it has been shown that the utility of type 3 agents is $EU_3 = 1.41$ in the post-intervention *LIME*. It is also easy to compute $EU_1 = b + 5.23$ and $EU_2 = 2.46$. Thus the policy results not only in a Pareto improvement over the *LIME*, but over the *LIE* as well.[15] This establishes that, in the presence of the suggested policy, restriction of competition in the issue of liabilities can produce Pareto improvements.

---

[14] Note that the government lends goods in the setting described in Section II, Part A. However, it does not matter if the government lends fiat currency instead, so long as loans must be repaid in money, and taxes and transfers are payable only in money.

[15] Type 1 utility increases because of the reduction in variance accompanying the policy. This was omitted in the approximation used for Proposition 1, which explains why type 1 agents experience a utility increase as opposed to the constant utility predicted by the proposition. The disparity arises because credit rationing in this example is not small.

## IV. Conclusions

For some time, credit rationing was thought to be the channel through which monetary policy operated. Considerable attention has been focused for this reason on the "micro foundations" of credit rationing. Among the reasons deduced for its existence is asymmetric information. The analysis here has adopted this approach in a framework sufficiently rich to allow for the presence of money and a discussion of the welfare implications of alternate policies. It was shown that when money is present, under extremely plausible circumstances (except for possibly the initial old generation), it is optimal to make liberal use of the discount window, or of government lending. Moreover, for some economies a policy where the government lends as much as desired at above market rates of interest is both *Pareto improving* and deflationary. Finally, in the presence of such a policy, only a limited set of factors under the control of the government influence private market rates of interest and credit volume.

Note that these results are not deducible in the absence of money. Consideration of a steady-state monetary equilibrium alters the analysis since the presence of money bounds below the rate of return on private debt. This is a forceful argument that if credit rationing is to provide a role for monetary policy, this policy must be analyzed in the context of the model generating the credit rationing. This paper seems to be the first to attempt such an analysis.

Moreover, it is also not the case that these results are deducible in the absence of asymmetric information. In particular, one could consider a world with full information where risky borrowers exercise monopsony power in credit markets. It is not the case that such an economy behaves like the one examined here. Specifically, it can be verified by considering the economy of examples 1–3 that the policy suggested is not welfare improving in a full-information setting. Thus the informational asymmetry is also essential to the analysis.

Finally, the analysis does *not* provide an irrelevance argument for further government policy actions. As an example, suppose that the rule governing the evolution of the money stock is amended to $M_t = (2/3)M_{t-1}$. Then the steady-state deflation rate becomes $S_{t+1}/S_t = 3/2$, and $\hat{R}$ rises for the example in the text.

In contrast to the discussion over the neutrality of money, then, policy analysis is fundamentally different in the presence of valued fiat money. Thus, if credit rationing operates and provides a role for monetary policy, money must be present in the model. If it is present, the result is an emphasis on discount window and government lending policies. It remains to be demonstrated if, in the presence of such policies, other monetary policies will be effective, or desirable.

## REFERENCES

**Barro, Robert J.**, "The Loan Market, Collateral, and Rates of Interest," *Journal of Money, Credit and Banking*, November 1976, *8*, 439–56.

**Diamond, Peter**, "Efficiency with Uncertain Supply," *Review of Economic Studies*, July 1980, *47*, 645–51.

**Fried, Joel and Howitt, Peter**, "Credit Rationing and Implicit Contract Theory," *Journal of Money, Credit and Banking*, August 1980, *12*, 471–87.

**Harris, Milton and Townsend, Robert M.**, "Resource Allocation under Asymmetric Information," *Econometrica*, January 1981, *49*, 33–64.

**Jaffee, Dwight M. and Russell, Thomas**, "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics*, November 1976, *90*, 651–66.

**Prescott, Edward C. and Townsend, Robert M.**, "Optima and Competitive Equilibria with Adverse Selection and Moral Hazard," Center for Economic Research, Department of Economics, University of Minnesota, June 1981.

**Rothschild, Michael and Stiglitz, Joseph**, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics*, November 1976, *90*, 629–49.

**Samuelson, Paul A.**, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *Journal* *of Political Economy*, December 1958, *66*, 467–82.

**Sargent, Thomas J. and Wallace, Neil,** "The Real Bills Doctrine vs. the Quantity Theory: A Reconsideration," Federal Reserve Bank of Minneapolis, January 1981.

**Smith, Bruce,** "Limited Information, Money, and Competitive Equilibrium," Federal Reserve Bank of Minneapolis, November 1982.

**Smith, Vernon L.,** "A Theory and Test of Credit Rationing: Some Generalizations," *American Economic Review*, June 1972, *62*, 477–83.

**Stiglitz, Joseph E. and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

**Wilson, Charles,** "A Model of Insurance Markets with Incomplete Information," *Journal of Economic Theory*, December 1977, *16*, 167–207.

# Effects of Interest Rates and Inflation on Aggregate Inventory Investment in the United States

*By* M. A. AKHTAR*

Until recently, econometric studies of inventory behavior have failed to uncover any significant effects of capital costs or interest rates on investment in inventories.[1] In fact, after many years of investigations, Michael Lovell concluded that "...the probability of obtaining an interest rate coefficient with negative sign is 50 percent" (1976, p. 400), and that inventories are not significantly influenced by capital costs. This conclusion, though perhaps justifiable in terms of the previous empirical studies, is difficult to reconcile with generally high inflation rates and rather large movements in real and nominal interest rates over the 1970's. Indeed, recent studies by F. Owen Irvine, Charles Lieberman, and Laura Rubin using time-series data over the 1960's and 1970's seem to indicate that capital costs may be important in determining the inventory behavior. However, many others including those by Alan Blinder (1981b), Martin Feldstein and Alan Auerbach (1976), and Louis Maccini and Robert Rossana (1981) continue to suggest that inventories are insulated from credit crunches or interest rate effects.

Almost all recent inventory studies have focused on the behavior of particular sectors or components of total inventories. It is therefore difficult to generalize the results at the national or macroeconomic level. Yet, as is well known, changes in aggregate inventory investment account for a large proportion of changes in Gross National Product ($GNP$) and its variance. For example, decreases in inventory investment accounted for two-thirds or more of the shrinkage in $GNP$ during the three recessions in the last decade—1969–70, 1974–75, and 1979–80. Perhaps, even more significantly, changes in inventory investment accounted for about one-third of the variance of changes in $GNP$ over the period 1959–79.[2]

Bearing in mind the significance of inventory movements for overall economic activity, this paper takes a fresh look at inventory carrying cost effects by focusing on aggregate inventory behavior. Section I outlines the basic estimation framework and Section II explains the empirical specification of the variables utilized in estimating the model. The third section discusses the results with particular focus on the effects of interest rates and inflation on aggregate inventory investment. The next section examines post-sample simulations based on estimates in Section III. The final section highlights the main findings and examines some possible explanations of the differences in the estimated carrying cost effects in this study vis-à-vis most earlier studies.

---

*Research Officer (Manager), International Research Department, Federal Reserve Bank of New York. An earlier version of this paper was presented in seminars at Columbia University and the Federal Reserve Bank of New York. I thank participants in those seminars, especially Phillip Cagan, William Dellalfar, Edward Frydl, and Lucas Papademos for comments and suggestions. I am also grateful to two anonymous referees for helpful criticisms and to Christopher Duffy and Andrew Gordon for competent research assistance. Neither these individuals nor the Federal Reserve Bank of New York share any responsibility for the views expressed in this paper.

[1] See David Savage and F. Owen Irvine (1981b) for recent brief surveys of the empirical literature on this subject. Many theoretical models have incorporated carrying cost effects on inventories; see, for example, Alan Blinder and Stanley Fischer, Charles Holt et al., and T. M. Whitin.

[2] Such calculations are not meant to suggest that autonomous inventory changes cause cyclical movements in economic activity, but they do indicate that inventories play a major role in propagating those movements. For a detailed analysis of the importance of inventory movements in business cycles, see Blinder (1981a,b), and Blinder and Fischer.

## I. Estimating Equation

The empirical analysis in this paper is based on the "stock adjustment" or the so-called "flexible accelerator" model. This model postulates that each firm has a desired level of inventories and that it attempts only a partial adjustment towards the desired level within any one period. Specifically,

$$(1)\quad INV_t - INV_{t-1} = \Delta INV_t$$
$$= \Delta INV_t^p + \Delta INV_t^u$$
$$= \alpha(INV^* - INV_{t-1}) + \Delta INV^u,$$

where

$INV_t$ = aggregate level of inventories on national income and product accounts (*NIPA*) basis, held at the end of period $t$;

$INV^*$ = aggregate level of inventories *desired* to be held at the end of period $t$;

$\Delta INV_t$ = change in aggregate inventories on *NIPA* basis over period $t$, which is divided into planned (or intended) component, $\Delta INV_t^p$, and unplanned (or unintended) component, $\Delta INV_t^u$;

$\alpha$ = "speed-of-adjustment" or stock adjustment coefficient, $0 < \alpha \leqslant 1$.

The unobservable desired level of inventories is hypothesized to be determined by expected demand (i.e., sales), carrying costs, and cyclical considerations:

$$(2)\quad INV_t^* = a_0 + a_1 DX_t^e + a_2 RS_t$$
$$+ a_3 P_t^e, + a_4 CU_t$$

where

$DX_t^e$ = expected aggregate sales of goods on *NIPA* basis in period $t$;

$RS_t$ = short-term interest rates in period $t$;

$P_t^e$ = expected price inflation in period $t$;

$CU_t$ = capacity utilization in period $t$;

and $a_1, a_3 > 0, a_2, a_4 < 0$.

The unplanned or unintended inventory investment will accumulate or be depleted in response to movements in capacity utilization and in actual demand relative to actual output. This component of inventory investment is specified as

$$(3)\quad \Delta INV_t^u = b_0 + b_1 CU_t + b_2(DX_t/Q_t),$$

where

$DX_t$ = aggregate sales of goods on *NIPA* basis in period $t$;

$Q_t$ = industrial output in period $t$;

and $b_1 > 0, b_2 < 0$.

Inserting equations (2) and (3) into (1), and adding a dummy variable yields

$$(4)\quad \Delta INV_t = c_0 + c_1 DX_t^e + c_2 RS_t + c_3 P_t^e$$
$$+ c_4 CU_t + c_5(DX_t/Q_t)$$
$$+ c_6 LDUM + \alpha INV_{t-1} + u,$$

where *LDUM* is a dummy variable, representing the 1980 general credit controls, equal to 1 during quarters 2 and 3, 1980, and 0 elsewhere; $c_0 = \alpha a_0 + b_0$; $c_i = \alpha a_i$, for $i = 1,2,3$; $c_4 = \alpha a_4 + b_1$; $c_5 = b_2$; and $u$ = a random disturbance term. Note that the sign of $c_4$ is ambiguous and depends on whether the planned or unplanned inventory behavior dominates capacity utilization.

## II. Specification of Variables

### A. *Expected Demand*

Nearly all studies of inventory behavior have postulated that expected sales are some function of past sales. Two widely used schemes for this purpose have been 1) to use the simple expectations model which assumes that expected sales are equal to sales of the last period, and 2) to specify expected sales as a distributed lag over sales in past periods. In this paper, both alternatives were tried in specifying $DX_t^e$ or expected quarterly sales of goods on a national income and product accounts basis. The first alternative is straightforward, that is, $DX_t^e = DX_{t-1}$. Under the second alternative, $DX_t^e$ was specified as an unconstrained second-degree polynomial on the preceding four quarters, or as a distributed lag estimated freely with the length of lag set at two or three quarters.

The simple model with $DX_t^e = DX_{t-1}$ seems to provide the best fit although the results based on the expected demand proxy with the polynomial lag structure are not significantly different. This is not surprising since the polynomial weight for $DX_{t-1}$ is 0.9 or higher. The estimates reported in this paper use the first alternative.

Since $DX^e$ is based entirely on past sales, it may not be a good proxy for capturing expectations of future sales. Therefore, in addition to $DX^e$, new orders for manufacturing goods as a proxy for expected sales was tried. (Note that manufacturing goods are a subset of $DX$.) The variable should be positively related to the desired level of inventories because increases in actual or planned output induced by additional new orders lead to increases in planned inventories of finished goods, and of raw materials and works-in-progress. However, in my empirical analysis, new orders for manufacturing goods did not turn out to be statistically significant, and they are therefore not included in the final reported estimates.

### B. *Carrying Costs*

The capital costs of financing inventories consist of two parts: 1) nominal carrying costs of inventories or the opportunity cost of funds tied up in financing inventory stocks; and 2) the anticipated rate of nominal capital gains associated with the price appreciation of goods held in inventory. The usual procedure is to treat the difference between these two elements ($RS - P^e$) as the appropriate measure of real carrying costs of inventories.

This type of real interest rate measure uses equal weights on nominal interest rates and expected inflation. However, several considerations point out that businesses may not view the coefficients of the two variables to be equal in size. In general, there is a good deal more uncertainty about $P^e$ than about $RS$. Therefore, as noted by Irvine (1981a), risk-averse firms are likely to weight $P^e$ less than $RS$. Similarly, Alan Blinder (1981b) has pointed that the nominal interest rate is the appropriate carrying cost for businesses following a first in,. first out (*FIFO*) pricing strategy—under which prices of goods in stock are not changed regardless of what happens to prices of newly produced goods. In other words, the weight on $P^e$ is zero since there is no price appreciation on goods held in inventory. Finally, the differences in tax treatment of interest income or expense and capital gains would lead to different after-tax real interest rates for an equal-sized change in $RS$ and $P^e$. For all these reasons, I use $RS$ and $P^e$ as separate variables in equation (4).[3]

The four-to-six month commercial paper rate in the current period seems to be a good measure of the opportunity cost of funds invested in inventories. However, it is more difficult to find a suitable measure of expected price inflation. There is no clear-cut case for any one price expectations model, or for any one type of price index. I considered three different schemes for price expectations using the producer price index and the *NIPA* based price deflator for inventory investment. Specifically, those schemes are

(*i*) static expectations, $P_t^e = P_t$;
(*ii*) adaptive expectations, $P_t^e = \phi P_t + (1-\phi)P_{t-1}$, where $0 < \phi < 1$;
(*iii*) rational expectations with perfect foresight, $P_t^e = P_{t+i}$, where $i \geqslant 1$.

The most plausible results are obtained with the simple adaptive expectations proxy using equal weights over the four quarters from period $t$-4 to period $t$. The producer price proxy from period $t$-5 to period $t$-1 also yields similar but statistically less significant results. Judged in terms of the standard error of the coefficient of the price variable and the overall regression fit, the producer price index yields considerably better results than the inventory deflator. My main results are based on the adaptive expectations proxy for producer prices over the most recent four quarters to period $t$. However, the empirical analysis also provides a comparison with certain other proxies for price expectations.

---

[3] There is no explicit consideration of taxes as a factor influencing carrying costs. Since there has been no significant change in the corporate income tax rate over the last fifteen years or so, not much is lost by treating the tax rate as a constant. Of course, as noted in the text, the differences in the tax treatment of interest income/expense and capital gains are accounted for by my use of $RS$ and $P^e$ as separate variables. See Blinder (1981b, pp. 483–84) for an analysis of this point.

### C. *Other Variables*

Equation (4) includes two other variables, $CU_t$ and $(DX_t/Q_t)$, which need some elaboration. The variable $CU_t$ is proxied by the Federal Reserve Board's manufacturing capacity utilization rate. As seen in equations (2) and (3), capacity utilization appears as an explanatory variable for both planned and unplanned changes in inventories. The relationship of $CU_t$ with the desired level of inventories is postulated to be negative as businesses tend to smooth out the pattern of production in order to avoid or minimize the loss of efficiency (and other costly changes) associated with rapid changes in the production process by adjusting the desired level of inventories. This means that at high-capacity utilization levels, businesses will tend to reduce inventory stock rather than to make production changes to meet current demand, while at low-capacity utilization levels, they will tend to build-up stocks to avoid further cuts in production.[4]

However, $CU_t$ is positively related to unintended changes in inventories, equation (3). The rationale is based on the assumption that over a long period, expected or potential output should be equal to expected demand. Thus increases in actual output relative to equilibrium output or demand would lead to unintended increases in inventory investment, *ceteris paribus*.

Since actual demand for goods often deviates from expectations, many previous studies have included some proxy for sales surprises. I have used the ratio of current sales of goods to current industrial output $(DX_t/Q_t)$ to capture this type of effect. This allows for sales surprises as well as for output surprises in the short run. Unintended inventory investment would accumulate (be depleted) when current sales are weaker (stronger) relative to current output. To the extent sales surprises are offset by output surprises, inventory investment may remain unchanged.

## III. Estimation Results

Equation (4), as specified above, was estimated with quarterly data for the sample period 1965–81, and for the subperiod 1972–81. The starting point of the full sample period has no particular significance except that it provides a substantial number of observations with similar quality data. The choice of the subperiod is based on the fact that the period since 1972 is marked with generally higher and more variable inflation and interest rates than the earlier part of the sample period. (The exact starting point for the subperiod is $Q1$, 1972 but there is no significant difference in the results if that point is shifted forward by one to three quarters.) Presumably, the combination of an upsurge in inflation, the 1973–74 oil price shock and the advent of generalized floating exchange rates may have led to a structural shift in 1972–73.

Given the inertia in adjustment of inventories, one would expect the errors to be serially correlated. This a priori expectation was confirmed by Durbin's $h$-statistic calculated for estimates of equation (4) based on the ordinary least squares ($OLS$) method.[5] As is well known, the combination of the lagged endogenous variable and serially correlated disturbances means that the $OLS$ estimates will not be consistent. There is no unique method of dealing with serial correlation when the lagged endogenous variables are present. I used two alternative techniques for obtaining consistent and more efficient estimates. First, equation (4) was estimated by instrument variables with correction for serial correlation under the Wallis method, as described by Edmond Malinvaud (pp. 557–59). Second, the Hildreth-Lu search procedure was utilized.[6] The two sets of

---

[4] See Paul Burrows for more details on the relationship of capacity utilization to production-smoothing behavior of firms. George Hay's model includes the lagged output level to account for the production-smoothing behavior. My estimates do not favor the inclusion of that variable. The lagged industrial output level used as a substitute for capacity utilization or as an additional variable yields statistically insignificant coefficients.

[5] See J. Johnston, pp. 312–13, for details of this test.

[6] For a comparison of the Hildreth-Lu procedure with other estimation procedures for autocorrelated errors, see G. S. Maddala (pp. 277–84). Note that, as demon-

TABLE 1—ESTIMATES OF AGGREGATE INVENTORY INVESTMENT
DEPENDENT VARIABLE: $\Delta INV_t$

| | 1965:III–1981:II | | 1972:I–1981:II | |
|---|---|---|---|---|
| Coefficients | Instrument Variables | Hildreth-Lu | Instrument Variables | Hildreth-Lu |
| Constant | 244.55 | 245.98 | 275.77 | 283.22 |
| | (5.74) | (5.80) | (4.74) | (5.09) |
| $DX^e$ | 0.30 | 0.30 | 0.31 | 0.32 |
| | (5.89) | (5.94) | (6.14) | (6.17) |
| $RS_t$ | −1.79 | −1.79 | −2.23 | −2.26 |
| | (7.33) | (7.44) | (7.36) | (8.04) |
| $P^e$ | 0.65 | 0.66 | 0.87 | 0.89 |
| | (3.61) | (3.68) | (4.13) | (4.58) |
| $CU_t$ | −0.26 | −0.27 | −0.38 | −0.42 |
| | (1.18) | (1.22) | (1.31) | (1.52) |
| $DX_t/Q_t$ | −1.92 | −1.93 | −2.27 | −2.32 |
| | (8.53) | (8.62) | (6.14) | (6.61) |
| $LDUM$ | −6.53 | −6.65 | −8.00 | −8.38 |
| | (2.51) | (2.58) | (3.32) | (3.74) |
| $INV_{t-1}$ | −0.60 | −0.61 | −0.58 | −0.59 |
| | (5.54) | (5.60) | (4.72) | (5.10) |
| $\bar{R}^2$ | 0.83 | 0.84 | 0.91 | 0.92 |
| Standard Error of the Equation | 3.25 | 3.25 | 3.13 | 3.10 |
| Durbin-Watson | 2.03 | 1.99 | 2.27 | 2.11 |
| $\rho$ | 0.02 | 0.01 | 0.17 | 0.09 |
| Estimated Elasticities of $INV^*$ with respect to | | | | |
| Expected Sales | 0.96 | 0.96 | 1.00 | 1.02 |
| Interest Rates | −0.089 | −0.089 | −0.104 | −0.105 |
| Expected Inflation | 0.026 | 0.026 | 0.045 | 0.046 |

*Notes:* The lagged inventory levels were treated as endogenous and exogenous variables lagged one or two periods were used as instruments; estimated $t$-values are reported in parentheses below the coefficients, those around or above 1.7 are significant at the 95 percent or higher confidence levels (one-tail test); $\rho$ is the final value of the autocorrelation parameter after several iterations.

estimates are almost identical. In this paper, I discuss estimates based on instrument variables technique. However, as a starting point, Table 1 reports the basic estimates from both procedures for the full sample period as well as for the subperiod.

Estimates in Table 1 indicate that all explanatory variables, except capacity utilization, are statistically very significant. This is true for both the entire sample period and the subperiod. However, the overall regression fit is closer for the subperiod than for the full period, which seems to be consistent with the notion of a shift in the underlying structure. This impression is confirmed by Chow test statistics calculated for estimates in Table 1. The calculated value of that test is 2.4 which rejects the hypothesis that observations since early 1972 follow the same structure as in the earlier period at the 5 percent level. But the null hypothesis is not rejected at the 1 percent level.

To the extent that there has been a shift, it seems to be associated with the behavior of interest rates, expected inflation, and the sales/output ratio variable. The coefficients of those variables are larger in absolute value in the subperiod than in the full sample period. By contrast, all regressions in Table 1 yield virtually identical coefficients for expected sales and lagged inventory stock. Both

---

strated by Roger Betancourt and Harry Kelejian, the Cochrane-Orcutt estimation procedure may be inconsistent when the lagged endogenous variables are present.

coefficients bear theoretically expected signs and are significantly different from zero at the 1 percent level.

The speed of adjustment—the coefficient of the lagged inventory stock—is estimated to be around 60 percent, that is, around three-fifths of the discrepancy between desired inventories and the lagged inventory stock is corrected within one quarter. This is a reasonably prompt speed of adjustment and effectively counters the objection by Feldstein-Auerbach and others that the flexible accelerator model is not an appropriate specification because the estimated adjustment coefficients turn out to be implausibly low.

Turning to the main focus of this paper, estimates in Table 1 indicate that, contrary to most previous studies, nominal interest rates and inflation have statistically significant effects on the behavior of inventories. The signs of the two coefficients are in accord with a priori expectations and test to be statistically different from zero at the 1 percent level. According to these results, a one percentage point rise in short-term interest rates leads to, *ceteris paribus*, a \$1.8–2.2 billion decline in real aggregate inventory investment, while a similar rise in expected inflation leads to a \$0.7–0.9 billion increase in inventory investment.

The coefficients of *desired* inventory stock with respect to the main arguments—expected sales, interest rates and expected inflation—in equation (2) can be obtained by dividing the estimated coefficients by the speed of adjustment, $\alpha$. The corresponding point-of-means elasticities based on those coefficients are shown in the last three rows of Table 1. The estimated interest rate elasticity is more than twice the absolute value of expected inflation elasticity. But, given that fluctuations in interest rates or capital costs tend to be frequently quite large, the estimated elasticities for both the interest rate and inflation imply very substantial effects on desired inventory levels.[7]

Three interrelated points need to be made about the estimated carrying cost effects. The first concerns the effects of nominal vs. real carrying costs on the inventory behavior. In my results, both influences appear to be important suggesting that businesses react to changes in nominal interest rates and not just to changes in real costs. If the real cost had been the appropriate variable, the coefficients on interest rates and price inflation would have been roughly equal with offsetting signs. In fact, the signs are offsetting, but the results imply a substantial *net* effect from nominal interest rates on inventory investment.

Results obtained by substituting $(RS_t - P_t^e)$ for $RS_t$ and $P_t^e$ as separate variables seem to support these observations. Consider the counterparts for the two instrument variables estimates in Table 1:

*Full Sample Period*

$$\Delta INV_t = \underset{(4.55)}{242.83} + \underset{(3.99)}{0.25DX_t^e}$$

$$\underset{(3.12)}{-0.78(RS_t - P_t^e)} \underset{(0.90)}{-0.25CU_t}$$

$$\underset{(6.25)}{-1.83(DX_t/Q_t)} \underset{(4.25)}{-0.56INV_{t-1}}$$

$$\underset{(1.34)}{-4.71LDUM} \quad \bar{R}^2 = 0.70;$$

$$SEE = 3.92;\ D\text{-}W = 1.90;\ \rho = 0.05.$$

*Truncated Sample Period*

$$\Delta INV_t = \underset{(2.36)}{212.83} + \underset{(3.24)}{0.28DX_t^e}$$

$$\underset{(1.94)}{-0.75(RS_t - P_t^e)} \underset{(0.13)}{-0.06CU_t}$$

$$\underset{(2.68)}{-1.55(DX_t/Q_t)} \underset{(3.42)}{-0.66INV_{t-1}}$$

$$\underset{(0.82)}{-3.48LDUM} \qquad \bar{R}^2 = 0.73;$$

$$SEE = 4.43;\ D\text{-}W = 1.96;\ \rho = 0.02.$$

---

[7] Irvine (1981b) reaches a similar conclusion on the capital cost effects on target retail inventory levels. However, unlike the present study, he estimated nearly identical point-of-means elasticities of desired inventory levels with respect to interest rates and inflation. Incidentally, the corresponding elasticities of inventory investment with respect to interest rates and expected inflation are quite large: $-1.6$ and 0.5 over the entire sample period while $-2.3$ and 1.0 over the truncated sample period, respectively.

TABLE 2—COMPARISON OF ESTIMATED REGRESSION COEFFICIENTS OF EXPECTED INFLATION PROXIES

| Estimation Period | $(PI_t/PI_{t-1})$ | $(PI_{t+4}/PI_t)$ (1) | $(PI_{t+4}/PI_t)$ (2) | $(PI_t/PI_{t-4})$ |
|---|---|---|---|---|
| Full Sample Period | | | | |
| Producer Price Index | 0.12 | 0.17 | 0.57 | 0.65 |
| | (0.84) | (0.72) | (1.50) | (3.61) |
| Inventory Deflator | 0.02 | −0.10 | 0.19 | 0.46 |
| | (0.12) | (0.68) | (0.68) | (2.47) |
| Truncated Sample Period | | | | |
| Producer Price Index | 0.11 | 0.65 | 0.17 | 0.87 |
| | (0.60) | (1.78) | (0.72) | (4.13) |
| Inventory Deflator | 0.04 | 0.22 | −0.11 | 0.65 |
| | (0.25) | (0.81) | (0.68) | (2.55) |

*Notes:* $PI$ = price index, either the producer price index or the inventory deflator; data sources are given in the Appendix; estimates in the second column under rational expectations proxy, $(PI_{t+4}/PI_t)$, are based on instrument variables technique suggested by B. T. McCallum, and use past values of inflation as instruments; estimated $t$-values are reported beneath the coefficients, those around or above 1.7 are significant at the 95 percent or higher confidence levels (one-tail test).

The coefficient of $(RS_t - P^e)$ is considerably more than one standard error smaller in absolute value and statistically less significant than the coefficient of $RS_t$ in the comparable estimates in Table 1. In the subperiod estimate, the real cost variable is not significant at the 99 percent confidence level. Moreover, the general properties of these estimates are inferior to those of the comparable estimates in Table 1; standard errors of estimates are substantially larger and $t$-statistics are lower for all coefficients.

The second point to be made about the estimated carrying-cost effects concerns the expected inflation variable. It was noted earlier that after considering several proxies for expected inflation, my main results utilize the simple adaptive expectations for producer prices with equal weights over the most recent four quarters. Other inflation proxies include static expectations under which $P_t^e = P_t$, adaptive expectations with declining or equal weights over the past two-to-six quarters, and rational expectations with perfect foresight, looking ahead one-to-four quarters. Table 2 reports the coefficients for three such proxies when they are substituted in estimating equation (4).

Of the three proxies, $P_t^e = PI_t/PI_{t-4}$ yields statistically superior results for both the producer price index and the inventory deflator. While the coefficients based on both indices are significantly different from zero, those based on the producer price proxy are statistically more significant over the full sample period as well as over the subperiod. This producer price specification is used in my main estimates. Note that other proxies for expected inflation turn out to be generally statistically insignificant, and in some cases bear the wrong sign.

The third point deals with the possibility of a significant shift in the estimated coefficients of interest and inflation rates. The absolute size of the two coefficients is a bit more than one standard error larger in the subperiod than in the entire sample period. The implied elasticities for both inventory investment and desired inventory stock are also larger in the subperiod. The notion of a shift in coefficients was formally examined by estimating the following regression for the full sample period:

$$\begin{aligned}\Delta INV_t = {} & 241.88 + \underset{(5.64)}{} \underset{(6.25)}{0.32 DX_t^e} - \underset{(2.96)}{1.33 RS_t} \\ & - \underset{(1.58)}{0.80 RS_t (DUM)} - \underset{(0.09)}{0.08 P_t^e} \\ & + \underset{(1.09)}{0.92 P_t^e (DUM)} - \underset{(8.85)}{1.95 (DX_t/Q_t)} \\ & - \underset{(1.17)}{0.26 CU_t} - \underset{(5.67)}{0.62 INV_{t-1}} \\ & - \underset{(2.98)}{7.67\, LDUM} \qquad \bar{R}^2 = 0.87;\end{aligned}$$

$$SEE = 3.23;\ D\text{-}W = 1.98;\ \rho = 0.01,$$

TABLE 3—DYNAMIC SIMULATION RESULTS[a]
($ BILLIONS)

| | | Prediction Errors (Annual Rate) from Simulations Based on Estimates for the Period: | | | |
|---|---|---|---|---|---|
| | Actual Aggregate Inventory Investment | 1965:III – 1981:II | 1965:III – 1979:IV | 1972:I – 1981:II | 1972:I – 1979:IV |
| 1980:I | −0.9 | | −0.80 | | 1.56 |
| II | 1.3 | | −6.91 | | −7.06 |
| III | −5.0 | | 1.67 | | −0.64 |
| IV | −7.2 | | −5.42 | | −1.11 |
| 1981:I | −1.4 | | −0.34 | | 2.71 |
| II | 10.8 | | 1.06 | | 4.25 |
| III | 14.9 | 18.83 | 17.89 | 19.42 | 19.83 |
| IV | 4.2 | 4.81 | 4.79 | 5.54 | 7.16 |
| 1982:I | −17.5 | −2.30 | −3.59 | 0.07 | 1.16 |
| *RMSPE*[b] | | | | | |
| 1980:I–1982:I | | – | 6.97 | – | 7.64 |
| 1980:I–1982:I[c] | | – | 3.61 | – | 4.06 |
| 1981:III–1982:I | | 11.30 | 10.89 | 11.66 | 12.19 |
| 1981:IV–1981:I | | 3.77 | 4.23 | 3.92 | 5.13 |

[a] Quarterly dynamic simulations were run by utilizing the estimated lagged stock of inventories (i.e., the stock of inventories was adjusted for each quarter by estimated changes in inventories) from the beginning of simulation periods.

[b] Root mean squared prediction error.

[c] Excluding 1981:III

where $(DUM)=1$ from 1972: I through 1981: II, and 0 elsewhere. In this regression, a test of the coefficients of $RS_t\,(DUM)$ and $P_t^e\,(DUM)$ would test the hypothesis that the coefficients of interest rates and inflation are equal in the long and the short period.

In fact, the estimates reported here are inconclusive. While they do not reject the equality hypothesis at the 5 percent level, they appear to be consistent with the possibility of a shift in the structure. The coefficient of $RS_t\,(DUM)$ is significantly different from zero at the 10 percent level (one-tail test). Also, the coefficient of $P_t^e$ is no longer significantly different from zero.

## IV. Simulation Results

This section briefly examines the postsample predictions. Such predictions are frequently viewed as a measure of the overall explanatory power (or forecasting ability) of estimates. Perhaps, more importantly, they are helpful in judging the extent of sensitivity or instability of results over the postsample period. Based on the estimates in Table 1 and their counterparts with 1979: IV as the cut-off point for the sample, several different dynamic simulation exercises were tried. The results from these exercises turned out to be quite similar. Table 3 reports simulation results from four such exercises.

The postsample simulations usually yield large prediction errors. While the results in Table 3 are no major exception in this respect, they appear to be good on conventional criteria. The root mean squared prediction errors (*RMSPE*) for simulations are relatively large, but in all cases this is mostly due to the unusually large prediction error in *Q*3, 1981. For the rest of the period, the *RMSPE*s are, in most cases, only 15–30 percent larger than the standard errors of estimates for the relevant regressions.

The *RMSPE*s for simulations based on the long sample period are somewhat smaller than those based on the truncated sample. However, in general, the differences do not appear to be significant and, from the forecasting perspective, it is difficult to make a strong case for one set of estimates over the other.

The prediction errors from various simulations with different termination points do not suggest any significant structural shift in the postsample period from 1979 year ending to early 1982. On the whole, the quarterly prediction errors from 1981:III to 1982:I based on estimates in Table 1 are roughly similar to those based on estimates with 1979:IV as the cut-off point for the sample period.[8] None of the simulations are able to capture the major buildup of inventories in 1981:III. The inventory accumulation in that quarter cannot be easily explained since it came in the face of a rise in interest rates, lower expected sales, and expectations of declining price inflation.

## V. Conclusions

The main finding of this study is that fluctuations in carrying costs—interest rates and inflation—have very significant effects on inventory investment. The estimated coefficients indicate that a one percentage point rise in short-term interest rates would decrease aggregate inventory investment by around $2 billion, and that a *pari passu* rise in expected inflation would offset less than half of that decline. Another important finding of this study is that the estimated adjustment coefficient between the desired inventory level and the lagged inventory stock indicates a reasonably prompt speed of adjustment.

The finding on inventory carrying cost effects is in sharp contrast to most of the earlier evidence on this subject. One obvious explanation would seem to be my choice of a more recent period for investigation as compared with most previous studies. The generally high inflation and interest rates over the late 1960's and the 1970's have implied substantially larger changes in capital costs than in the earlier period which was characterized by relatively low inflation and interest rates. This also means that interest costs as a proportion of total carrying costs (other costs include storage, deterioration and obsolescence) have increased over the 1970's so that they would be expected to have larger effects on inventory investment. To some extent, this is evident in my estimates over the truncated sample period, 1972–81. As noted in the introduction, some studies using recent time-series data arrive at results which are generally consistent with my finding on carrying costs. However, others using similar data period reach the opposite conclusion.

Another explanation for the differences in findings on carryings costs seems to be related to the specification of interest rates and inflation. The present analysis provides an explicit treatment of the two carrying cost components and utilizes an expectations proxy for price inflation. By contrast, most earlier studies have used either a "real" interest rate variable, or have combined nominal interest rates with actual current price inflation. As shown earlier, both of those alternatives appear to be less appropriate than the specifications used herein. However, even the real interest rate specification appears to be significant in my estimates, as it is in a recent study of the retail inventory behavior by Irvine.

The specification of the unintended or unplanned inventory investment may also have played a role in capturing the carrying cost effects. The behavior of actual inventories does not always closely follow the behavior of planned inventory levels and the influence of interest rates on planned inventories can be swamped by unintended changes arising from unexpected demand or output conditions. Unlike most earlier studies, I have utilized the ratio of actual current sales to actual current output together with capacity utilization to account for unintended inventory investment. This specification seems to be helpful in clarifying the effects of carrying costs, although those effects remain statistically significant even if unplanned inventory investment is specified in terms of sales prediction errors by businesses as is done by many earlier investigators.

Finally, one major policy implication of my results on inventory carrying costs is worth mentioning. Until recently, economists have widely believed that inventories are not a significant channel for transmission of monetary influences to the real sector. The findings in this study, by contrast, suggest that monetary policy actions which alter short-term interest rates have quantitatively significant effects on aggregate inventories, and thus add to the rather scarce literature confirming the significance of that channel of monetary policy influence on the economy.

---

[8] These results are consistent with the fact that the estimated coefficients of the main variables are quite insensitive to various termination points from 1979:IV to 1981:II.

## APPENDIX: DATA SOURCES

All data series are from *Survey of Current Business* published by the U.S. Department of Commerce, and are maintained in CITIBASE (Citibank economic database). The empirical analysis utilized seasonally adjusted series, expressed in 1972 prices, from data tapes as of April 1982.

## REFERENCES

**Betancourt, Roger and Kelejian, Harry,** "Lagged Endogenous Variables and the Cochrane-Orcutt Procedure," *Econometrica*, July 1981, *49*, 1073–78.

**Blinder, Alan S.,** (1981a) "Inventories and the Structure of Macro Models," *American Economic Review Proceedings*, May 1981, *71*, 11–16.

______, (1981b) "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, *2*: 1981, 443–505.

______ **and Fischer, Stanley,** "Inventories, Rational Expectations, and the Business Cycle," *Journal of Monetary Economics*, November 1981, *8*, 277–304.

**Burrows, Paul,** "Explanatory and Forecasting Models of Inventory Investment in Britain," *Applied Economics*, December 1971, *3*, 275–89.

**Feldstein, Martin S. and Auerbach, Alan,** "Inventory Behavior and Durable-Goods Manufacturing: The Target-Adjustment Model," *Brookings Papers on Economic Activity*, *2*: 1976, 351–96.

**Hay, George A.,** "Adjustment Costs and Flexible Accelerator," *Quarterly Journal of Economics*, February 1970, *84*, 140–43.

**Holt, Charles C. et al.,** *Planning Production, Inventories, and Work Force*, Englewood Cliffs: Prentice Hall, 1960.

**Irvine, F. Owen, Jr.,** (1981a) "Merchant Wholesale Inventory Investment and the Cost of Capital," *American Economic Review Proceedings*, May 1981, *71*, 23–29.

______, (1981b) "Retail Inventory Investment and the Cost of Capital," *American Economic Review*, September 1981, *71*, 633–48.

**Johnston, J.,** *Econometric Methods*, 2d edition, New York: McGraw-Hill, 1972.

**Lieberman, Charles,** "Inventory Demand and Cost of Capital Effects," *Review of Economics and Statistics*, August 1980, *62*, 348–56.

**Lovell, Michael C.,** "Comments" on "Inventory Behavior and Durable-Goods Manufacturing: The Target-Adjustment Model," *Brooking Papers on Economic Activity*, *2*: 1976, 397–408.

**Maccini, Louis J., and Rossana, Robert J.,** "Investment in Finished Goods Inventories: An Analysis of Adjustment Speeds," *American Economic Review Proceedings*, May 1981, *71*, 17–22.

**Maddala, G. S.,** *Econometrics*, New York: McGraw-Hill, 1977.

**Malinvaud, Edmond,** *Statistical Methods of Econometrics*, New York: North-Holland, 1980.

**McCallum, B. T.,** "Rational Expectations and the Estimation of Econometric Models: An Alternative Procedure," *International Economic Review*, June 1976, *17*, 484–90.

**Rubin, Laura S.,** "Aggregate Inventory Behavior: Response to Uncertainty and Interest Rates," *Journal of Post Keynesian Economics*, Winter 1979/80, *2*, 201–11.

**Savage, David,** "The Channels of Monetary Influence: A Survey of the Empirical Evidence," *National Institute Economic Review*, February 1978, *83*, 73–90.

**Whitin, T. M.,** *The Theory of Inventory Management*, Princeton: Princeton University Press, 1953.

**U.S. Department of Commerce, Bureau of Economic Analysis,** *Survey of Current Business*, Washington: USGPO, various issues.

# Competition and Unanimity Revisited

*By* LOUIS MAKOWSKI*

Some recent developments in the theory of competitive stock markets are discussed and illustrated in this paper, using simple examples. These developments challenge the central role played by "spanning" in the literature dealing with unanimity among shareholders. Harry DeAngelo's 1981 article gives a clear account of some popular views about the central importance of spanning, and is used to focus my presentation.

In DeAngelo's article, there are three basic claims:

(a) If all initial shareholders in a firm have perfectly competitive conjectures and there is spanning, net market value maximization will be unanimously favored by initial shareholders as the objective of the firm whether or not there are short sales of the firm's shares (perfect competition plus spanning imply unanimity).

(b) Perfectly competitive conjectures only make sense when there is spanning (perfect competition implies spanning).

(c) For perfectly competitive conjectures to be *correct*, there must not only be spanning, but also constant returns to scale, and hence zero profits (profits are inconsistent with perfect competition).

In this paper, I show that claim (a) can be true even without spanning. In particular,

(i) If all initial shareholders in a firm have correct and perfectly competitive conjectures, then net market value maximization will be unanimously favored by them as the objective of the firm, at least if there are no short sales of the firm's shares by initial shareholders (perfect competition, with or without spanning, implies unanimity).

(ii) Claim (b) above is false; that is, perfect competition can exist without spanning.

(iii) Claim (c) is false, whether or not there is spanning; that is, positive profits and perfect competition are perfectly consistent.

Since the assumption of spanning is tantamount to the assumption of complete markets (for example, if there is free disposal plus spanning then markets are complete (D. M. Kreps, 1979), assertion (i) significantly extends the scope of the unanimity result. In particular, it extends the result to environments with incomplete markets in which firms market unique securities. Assertion (ii) shows such environments exist, and (iii) shows they can be nontrivial, that is, ones in which shareholders care about the firm's objective beyond whether or not the objective implies nonnegative profits.

Assertions (i)–(iii) are easily demonstrated once it is realized that the old textbook maxim—for perfect competition there must be many buyers *and sellers* of a commodity (for example, a particular type of security)—is simply false. While many buyers and sellers may be sufficient for perfect competition, it is not necessary. Indeed,

(iv) There may be only *one* seller of a commodity—*without* any good substitutes—yet that seller will face a perfectly elastic demand if he is small relative to his market.

The import of (iv) is that spanning is basically irrelevant for perfect competition;[1]

---

*Cambridge University. The idea for this paper germinated after a discussion with Oliver Hart, and most of the ideas expressed are also due to Hart. But some of the perfect competition ideas are due to Joe Ostroy. The Appendix was motivated by a discussion with Derek Atkins. I have also benefited from comments by Robert Clower, Joe Ostroy, and John Riley. Research for the paper was supported by a grant from the U.K. Social Science Research Council.

[1] Observation (iv) is due to Oliver Hart. The idea appears in his 1979a article, and is further developed in his 1979b and 1980 articles. A complementary development, in the context of finite rather than large economies, appears in my 1980b and forthcoming articles. The link between these two lines is Joe Ostroy's "no surplus" characterization of perfect competition, for example, his 1980 and 1981 papers. He shows that the no surplus property characterizes perfectly competitive economies, whether they are finite or large, although it is more likely to hold in large economies. Since finite examples are generally easier to understand, this allows for a simpler exposition of perfect competition theory and its consequences.

all that is required is that any seller's supply of his commodity is small relative to the demand he faces.

The organization of the sequel may now be summarized. In Section I, I present some simple examples which develop the intuitions underlying the paper. The examples also establish the validity of assertions (ii) and (iii) above. Section II then proves my basic assertion (i), via a slight modification of DeAngelo's proof of claim (a) above. The proof should be easy to follow once Example 4 in Section I below is understood, since the assumptions and line of argument used in the proof only involve an abstract restatement of the principles underlying the example. Alternatively expressed, the example can serve as a guide for walking through the proof. Section III then illustrates the importance of the no short sales assumption in the statement of (i) above. Finally, in the Appendix, an alternate reformulation of my argument is presented.

A note on terminology: throughout a perfectly competitive seller is synonomous with one who faces a perfectly elastic demand for his product. Also, a seller with monopoly power is equivalent to one who faces a strictly downward-sloping demand curve for his product.

## I. Some Examples

The examples that follow represent the heart of this paper. Example 1 is preliminary and illustrates that perfect competition and positive profits are perfectly consistent (assertion (iii) above). The intuition underlying (iii) is that under perfect competition, profits reflect the economic rent on a firm's superior technology. They measure the "marginal product" of the firm's technology (see my 1980a article for a formal demonstration). All subsequent examples of perfect competition will also involve positive profits.

Example 2 illustrates the basic idea underlying (iv) above and the rest of the paper, including all subsequent examples. It is intended to show that the belief that for perfect competition there must exist *many* sellers of a good is just not true: many sellers may be sufficient for perfect competition, but it is not necessary. Once this premise is given up, indeed is substituted by the more subtle thought that in some cases having only one seller, and *no* seller of good substitutes, may be consistent with perfect competition, the rest of the paper will be easy to understand. The subsequent examples place the idea of Example 2 into a stock market context.

A note on terminology: spanning means that any one seller cannot change the menu of goods being marketed. In a stock market context, this means that no firm can change the set of available securities. In Example 2 below and all subsequent examples, there will *not* be spanning because the unique seller of a good *can* effect its availability. Yet in these examples, the unique seller will still be a perfect competitor—face a perfectly elastic demand for his good—as will become clear shortly.

One final note on the examples may be useful before I proceed. The examples that follow all involve only a few individuals —they are finite examples—yet they will be examples involving perfect competition. This reflects some manipulation: finite examples with perfectly elastic demands *have* to be carefully constructed because they are events which occur on sets of measure zero. The reader may thus be tempted to dismiss the examples as mere curioso. This would be a mistake because the examples illustrate *general principles* of perfect competition (for example, assertions (i)–(iv) above) that apply to both finite *and* infinite economies. And while perfect competition, with or without unique sellers, is a rare event in finite economies, it is a generic event in large economies. (The intuition of why what is extremely unlikely with small numbers becomes extremely likely with large numbers is developed further below. See especially the discussion at the end of Example 3.) The advantage of the finite examples is that they are easier to understand because of the simpler mathematics involved; hence the general *principles* of perfect competition are much more plain to see.[2]

---

[2] There is another way of expressing the above that may be more familiar to some readers: While my finite examples of perfect competition without spanning are

### A. *Example* 1: *Perfect Competition with Profits*

Suppose there are just three potential producers of a good $X$. Firm 1 can produce at most 4 units of $X$, at a cost of \$1 per unit. Firms 2 and 3 can produce any number of $X$'s, at a cost of \$2 per unit. Finally, suppose the market demand for $X$ is $D_X = 12 - P_X$. Then the competitive equilibrium price of $X$'s will be \$2, with 10 units sold: 4 by firm 1 and 6 by firms 2 and 3. (See Figures 1A and 1B.)

While firm 1 earns a profit of \$4 (really a rent on its superior technology), it is a perfect competitor: if it tried to charge more than \$2 per $X$, no one would buy from it, switching their demands instead to firms 2 and 3; and the market for $X$'s would still clear at \$2, without firm 1's production. Also note that firm 1 *contributes no surplus to any of its customers*: they can always go elsewhere and get the same consumers' surplus. This illustrates a general principle:

(v) A firm faces a perfectly elastic demand for its product if and only if it contributes no surplus to any of its buyers (no surplus characterization of perfect competition).

Note that (v) does not say under perfect competition buyers get no surplus from purchasing any given commodity; rather, it says they get no surplus from purchasing it from *any one given seller*. Of course, if some agent is the only seller of a commodity in the economy—there are no perfect substitutes, so no spanning—the two statements collapse into one. (For a general proof of this no surplus property, see my 1980a article. Note the connection between (iv) and (v): (iv) provides a sufficient environmental condition for perfect competition, while (v) provides a characterization of perfect competition, so, a necessary and sufficient condition.) The example illustrates that profits are not inconsistent with perfect competition (compare claim (c) above).

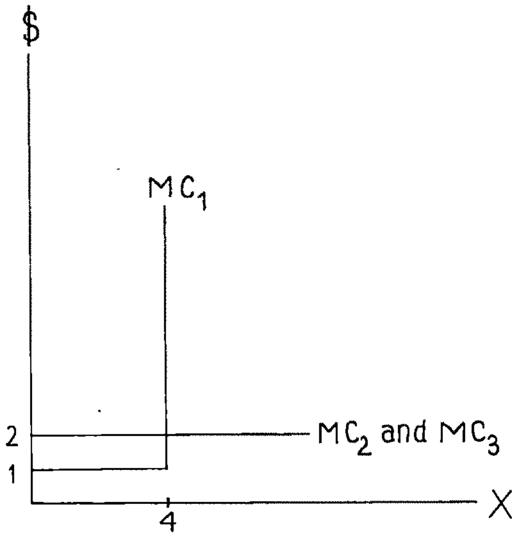


FIGURE 1A. THE FIRMS' COSTS

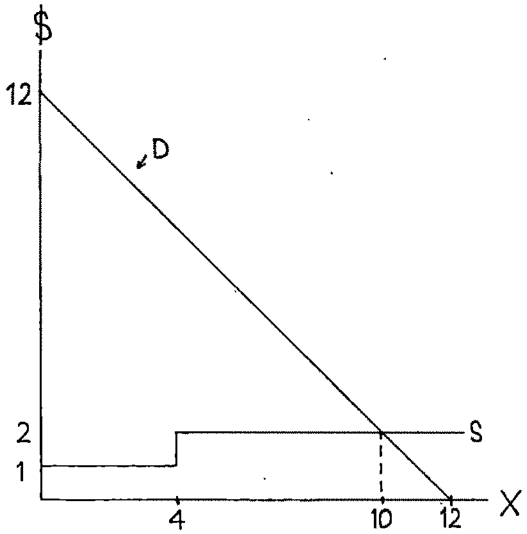


FIGURE 1B. THE MARKET EQUILIBRIUM IN EXAMPLE 1

As it stands, this is a partial equilibrium example, but it can easily be transformed into a general equilibrium one: Suppose there are ten individuals in the economy, each with a utility function

$$(1) \qquad U_i = 12X_i - 5X_i^2 - L_i,$$

where $L_i$ is individual $i$'s labor supply. Each individual can supply up to 10 hours of labor

---

certainly not "preference free," they become preference free "in the limit," i.e., in the large numbers case. (Again see the discussion ending Example 3 below.)

and has no endowment of $X$'s. Firm 1 can produce at most 4 units of $X$ using 1 hour of labor per unit; it is owned by individuals 1 and 2 in equal share. Firms 2 and 3 can produce $X$'s at constant returns to scale using 2 hours of labor per $X$. If we let labor be the numeraire, that is, the wage rate is \$1 per hour, then it is easy to check that in general equilibrium the cost, demand, and supply curves will be as in Figure 1, with each individual consuming $1X$ and each individual except 1 and 2 supplying 2 hours of labor; individuals 1 and 2 don't work, but finance their consumption via firm 1's distributed profits (\$2 of profit go to each). As before, profits are consistent with perfect competition; for example, if firm 1 tried to charge more than \$2 per $X$, no one would buy from it, switching their demands instead to firms 2 and 3. Also all prices would remain unchanged; that is, the markets for $X$ and labor would still clear at \$2 and \$1, respectively, without firm 1's production.

In this example there is spanning since, even if firm 1 did not produce $X$'s, the $X$'s would be available from other firms. In the following examples, we will not have spanning, but we will still have perfect competition.

### B. *Example* 2: *Perfect Competition without Spanning*

Suppose there are only three antique cars of a particular type available in the economy, all initially owned by one individual, $A$. Suppose each individual in the economy is willing to pay at most \$10,000 to own one such car, \$8,000 each to own two, \$6,000 each to own three, and so forth. The market demand $A$ faces is then as pictured in Figure 2A, where $I$ is the number of individuals in the economy. Note if $I < 3$ then in his effective range (3 or less cars) $A$ would face a downward-sloping demand curve. But if $I > 3$ then he faces a perfectly elastic demand as far as he is concerned, as illustrated in Figure 2A.

With $I > 3$, the competitive equilibrium price of cars would be \$10,000: If any individual refused to pay this sum to $A$, some other potential buyer—who does not have a car—would bid the car away from that individual. Note again that the no surplus property is satisfied: no buyer of cars gets any surplus in buying them from $A$. However, now there is no spanning since $A$ is the sole seller of antique cars.

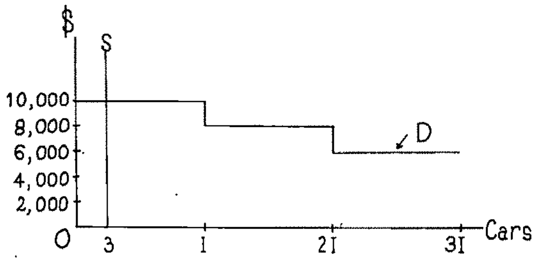


FIGURE 2A. MARKET EQUILIBRIUM FOR ANTIQUE CARS

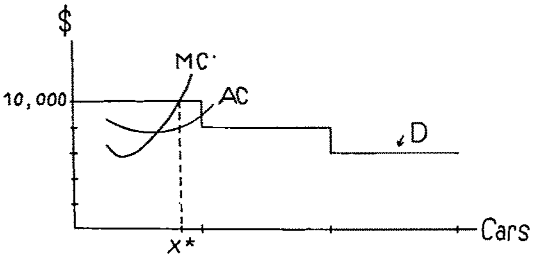


FIGURE 2B. MARKET EQUILIBRIUM FOR SPECIAL, PRODUCED CARS

What makes this example work is *not* that cars are in inelastic supply, only that $A$ is *small relative to his market*. To see this, suppose $A$ is selling not antique cars but a special type car that only he knows how to produce. And suppose he has convex cost curves as pictured in Figure 2B. Again $A$ faces a perfectly elastic demand in his effective range and thus is a perfect competitor. He will produce $x^*$ cars, and sell them at a price of \$10,000 each.

It should be apparent from the above that the indivisibility of cars is also not crucial to the example. The next example, a variant of Example 1, should make this clearer yet.

A warning against possible confusion: DeAngelo argues that without spanning, a seller *cannot* be a perfect competitor because he can dictate a finite price for his good (if made available) and an infinite price (if made unavailable) (see his article, p. 22). This argument is misleading because it is based on an implicit *pre*sumption, or convention. Even a *unique* seller of a good, given that he is a perfect competitor, will not be able to essentially affect the price of his good. Imagine a downward-sloping market demand curve for a good, which cuts the ordinate at some price $\bar{p}$ (for example, at \$10,000 in Figure 2A). This price is obviously the maximum of all potential buyers' reservation prices for the good. So, saying that all buyers' "perceived price" for the good is in the interval $[\bar{p}, \infty]$ may be analytically thought of as a way of saying the good isn't really available: at any price in $[\bar{p}, \infty]$ no one wants to buy, or is at best indifferent at $\bar{p}$. Which point in the interval one chooses is basically a matter of convention. My convention is to say that when the good isn't available, the perceived price is $\bar{p}$, the maximum reservation price; DeAngelo's convention is to say it is infinity. The advantage of the former convention is that *if* there is a unique seller of the good who happens to face a perfectly elastic demand for the good at $\bar{p}$—so he can get $\bar{p}$ for all units of the good he wants to sell, then whether or not he sells the good, its price won't change. This convention is useful in that the theory of perfect competition, that a perfect competitor cannot affect prices, is robust to the case of a single seller. The same *principles* apply if there is one seller of a good or many (for example, assertions (i)–(v), above). The convention helps highlight this unity.

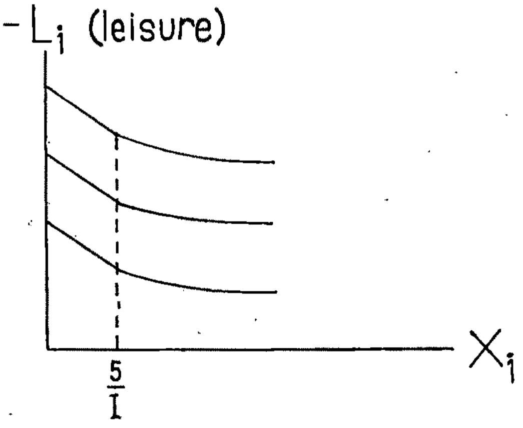


FIGURE 3A. AN INDIVIDUAL'S INDIFFERENCE MAP

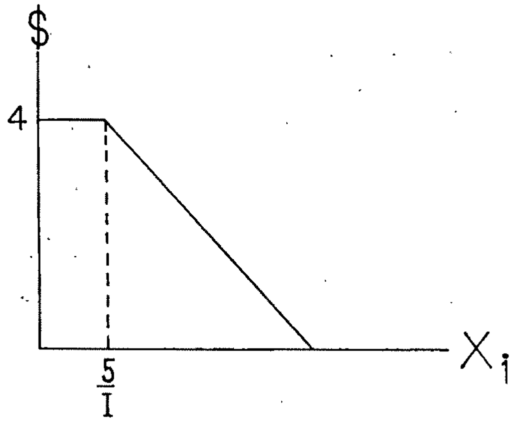


FIGURE 3B. AN INDIVIDUAL'S DEMAND FOR $X$'S

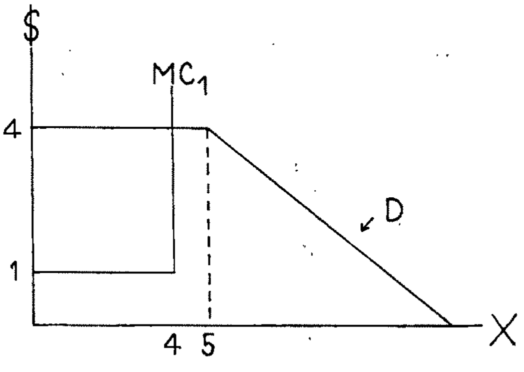


FIGURE 3C. THE MARKET EQUILIBRIUM IN EXAMPLE 3

### C. *Example* 3: *Perfect Competition with Profits and No Spanning*

As in Example 1, there are just two commodities, $X$ and $L$, and firm 1 can produce up to 4 units of $X$ using 1 hour of labor per $X$. But now firms 2 and 3 do not exist, so, firm 1 is the only potential producer of $X$'s.

Let there be $I$ individuals in the economy, each with the utility function

$$(2) \qquad \bar{U}_i = 4X_i - L_i \qquad \text{if } X_i \leqq 5/I;$$

$$\bar{U}_i = 4X_i + (5/I)X_i - \tfrac{1}{2}X_i^2 - L_i - \tfrac{1}{2}(5/I)^2 \qquad \text{if } X_i > 5/I,$$

where $I$ is a parameter depending on the size of the economy. Note as $I$ becomes large, $\overline{U}_i$ approaches $4X_i - \frac{1}{2}X_i^2 - L_i$ since $5/I$ approaches zero. A typical individual's indifference curves between $X$ and $-L$ (leisure) are illustrated in Figure 3. Note there is vertical parallelism, so all indifference curves have the same shape. They are convex, but begin with a linear segment until $(5/I)X$'s are consumed.

The importance of this linear segment is to insure firm 1 faces a perfectly elastic demand for $X$'s. In particular, it is easy to check that each individual's demand for $X$'s is as illustrated in Figure 3B (again $L$ is the numeraire, so the wage rate is \$1). Consequently, the market demand is as illustrated in Figure 3C. And in a competitive equilibrium each individual consumes $4/I$ units of $X$ and each individual except 1 and 2 works $16/I$ hours. Individuals 1 and 2 do not work, but even hire $6-16/I$ units of labor (servant services) out of firm 1's distributed profits (\$6 to each).

As in the previous examples, observe that the no surplus characterization of perfectly elastic demand holds: no buyer gets any consumer's surplus from purchasing $X$'s from firm 1. And profits are consistent with perfect competition. But again there is no spanning.

Also note that as $I$ gets larger and larger, the linear segment in each individual's indifference curves—needed to insure firm 1 is small *relative to his market*—gets smaller and smaller. In the limit, as $I$ approaches infinity, the indifference curves approach a smooth, *strictly* convex shape. This illustrates another general principle:

(vi) In finite economies, there must be some linearity in the production possibility frontier or aggregate net trade set to insure perfect competition. But in large economies —with many *buyers* but *not* necessarily many sellers of each commodity—differentiability or smoothness replaces linearity as sufficient for perfect competition.

This is intuitive since a differentiable function or smooth curve is locally linear to a first-order approximation. And it is consistent with the (correct) intuition that in large economies perfect competition is significantly more likely than in finite ones. This is because in large economies all buyers and sellers are almost automatically, by construction small relative to their markets (compare assertion (iv) above; for general proofs of (vi), see Ostroy, 1980, 1981, or my 1980c paper).

I turn next to a stock market example, illustrating assertions (i)–(iii).

### D. *Example* 4: *Perfect Competition in a Stock Market Economy with Profits and No Spanning*

This is a slight variant of the preceding example: suppose there are two periods, no uncertainty, and just two consumption goods, $Y$ and $X$; where $Y$ represents the first-period consumption good and $X$ the second-period consumption good. Again there are $I$ individuals in the economy. Each has the utility function

$$(3)\qquad \begin{aligned} V_i &= 4X_i + Y_i && \text{if } X_i \leqq 5/I; \\ V_i &= 4X_i + (5/I)X_i - \tfrac{1}{2}X_i^2 \\ &\quad + Y_i - \tfrac{1}{2}(5/I)^2 && \text{if } X_i > 5/I. \end{aligned}$$

Each is endowed with $10Y$'s and no $X$'s.

Again, there is only one firm in the economy. It can produce up to 4 units of $X$ using 1 unit of $Y$ as input per $X$ produced. The firm is initially owned by individuals 1 and 2, who hold one-half share each. Also, there are no bonds or futures markets. So, individuals can only transfer purchasing power between periods by transacting on the stock market (i.e., by buying and selling shares in the firm).

Letting $Y$ be the numeraire (so its price is \$1 per unit), a typical individual's demand for $X$'s is illustrated in Figure 3B, and the market demand is illustrated in Figure 3C. Consequently, a security promising to pay $x$ units of $X$ in period 2 would sell for \$$4x$ provided the number of such securities supplied, $n(x)$, satisfies $n(x)x < 5$; for example, if there were $n$ shares marketed by the firm, each entitling its holder to $1/n$th of the firm's second-period return (initial owners paying for input costs), the price of such a share would be \$$4X_f/n$ (where $X_f$ is the

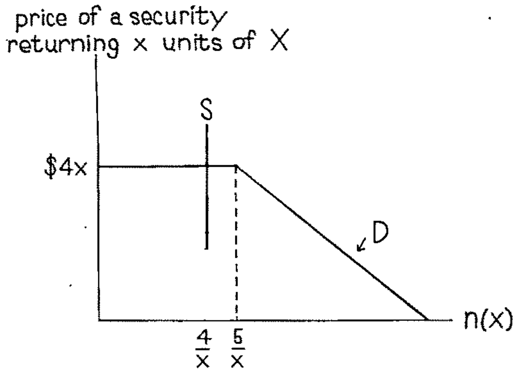


FIGURE 4. THE SECURITY MARKET EQUILIBRIUM

firm's second period output). (See Figure 4.) Note that since the firm can produce at most $4X$'s, $X_f \leqq 4$; so, $n(x)x \leqq 4$. Alternatively expressed, in its effective range the firm faces a perfectly elastic demand for its shares.

Clearly, the initial owners would like the firm to maximize its net market value—which is exactly analogous to maximizing profits in the preceding example. That is, the owners unanimously want it to produce $4X$'s, so each share will sell for \$16 (assuming the firm does not issue any new shares).[3] Then, in a competitive equilibrium each individual consumes $4/I$ units of $X$, costing $\$16/I$. Thus, each individual except 1 and 2 consumes $10-16/I$ units of $Y$, while 1 and 2 consume $16-16/I$ units of $Y$ each. Note if $Y=-L$, the consumption is the same as in the preceding example. But now in an equilibrium each individual holds $1/I$ shares, so individuals 1 and 2 sell $1/2I$ shares each to individuals $3,4,\dots,I$. Individuals 1 and 2 thus earn a total of $\$16(I-2)/2I$ each from sale of shares; or, net of input costs, $\$6-16/I$ each.

As in Example 3, notice that no buyer obtains any surplus from buying shares in the firm: the no surplus characterization of perfectly elastic demands is again illustrated. Also, perfect competition implies unanimity *even without spanning*: the firm's shares have no substitutes in this example. Finally, notice that initial shareholders are not just indifferent about the firm's objective: if the firm produced less than $4X$'s, its net market value would fall and all initial shareholders would be strictly worse off, although no *buyer* of shares would care since the price of the firm's shares would remain at the buyers' reservation price.

Thus, the example establishes the validity of my assertions (ii)–(iii), and illustrates (i). I am now ready to demonstrate the general validity of (i). Afterwards, the importance of the no short sales requirement will be discussed.[4]

## II. Perfect Competition and Unanimity

I shall now demonstrate assertion (i) above. To proceed, simply delete DeAngelo's Assumption 8 (p. 22), and replace his Assumptions 4 and 9 with the following.

ASSUMPTION 4*: (*No short sales of unique securities by initial shareholders.*) If a firm can change the availability of any pattern of state claims (i.e., any given type of security) by altering its production or capital structure decisions, then no initial shareholder would go short in that security. That is, either short selling of that security is not possible or no initial shareholder would *want* to go short in the security.[5]

ASSUMPTION 9*: (*No surplus characterization of perfectly elastic demand for shares.*) The production and capital structure decisions of any given firm do not affect the market price of current consumption goods or securities marketed by other firms. And, if these decisions make available (or make unavailable) any pattern of state claims (i.e., any given type of security), the market price of that security would be (was) sufficiently high so that no buyer of the security gains (loses) any consumer's surplus from the firm's decisions.

---

[3] It is easy to extend the example to allow the firm to finance investment by issuing new shares. As should be obvious from Figure 4, a Miller-Modigliani-type irrelevance of financing result would then emerge.

[4] Examples of perfect competition with uncertainty can easily be constructed along the lines of Example 4. For one such example see my forthcoming article. Other examples, including one similar to Example 2, appear in Hart (1979a).

[5] The reader may, as a first approximation to Assumption 4*, simply assume that if a firm's shares are not spanned by existing securities then short selling of its shares is not possible. This is strictly stronger than Assumption 4*, but also simpler. Hence, it will be easier to keep in mind.

The demonstration of (i) now simply follows the lines of DeAngelo's argument. As a result of Assumptions 4* and 9*, in a competitive market a firm's decisions can affect individuals' consumption opportunities only through changes in personal wealth. The set of attainable current consumption goods and their prices are fixed. While a firm may affect the set of available securities, those whose availability may change with a firm's decisions will never be available at a price below any potential buyer's reservation price, so their availability or unavailability does not matter to any potential buyer; and given Assumption 4*, it also does not matter to any potential short seller among the initial shareholders. But, for any initial shareholder,

> more wealth results in unambiguously greater consumption opportunities so that optimal corporate decisions are those which maximize personal wealth.... Now, a given firm's decisions are perceived as affecting an individual's wealth only through their impact on his proceeds from the t = 0 net value of that firm (Assumption 7 implies no cash flow impact on other firms and Assumption 9[*] implies no perceived valuation impact). And since every original owner shares proportionately in net firm value (by Assumption 6), maximizing net firm value is unanimously supported because it simultaneously maximizes the personal wealth and therefore the consumption opportunities of every owner.
>
> [DeAngelo, p. 23]

This completes the demonstration of (i). Note that if a firm cannot change the availability of any given type of security then, as in DeAngelo, it does not matter to the unanimity result whether or not the security may be sold short (provided any short selling is limited by a constraint against personal default).

Also note that, as in DeAngelo, if old financing is protected by an nonexpropriation clause, then the unanimity result can easily be extended to include "*ex post* shareholders." So there is no potential conflict between *ex ante* and *ex post* shareholders' wishes. Finally, note that I have actually established a stronger result than (i), since I have also shown that, under perfect competition, if one is not a shareholder in the firm then one will be indifferent about the firm's production and capital structure decision, provided one does not want to go short in the firm's shares.

## III. Short Selling and Unanimity

To understand the role of the no short selling Assumption 4*, it is best to consider another simple example.

### A. *Example* 5: *Perfect Competition without Unanimity*

As in Example 4, there are two periods, no uncertainty, and just two consumption goods, $Y$ and $X$. Also, there are no bonds or futures markets, and only one firm. Thus, individuals can only transfer purchasing power between periods by buying and selling shares in the firm.

The firm possesses only one possible production activity, to produce $1X$ using $6Y$'s as input (i.e., its production possibility set contains only two points $(-6,1)$ and $(0,0)$). And the firm is initially owned by individuals 1 and 2, who hold one-half share each.

Suppose now that individuals 1 and 2 only like consumption in period one, that is, each has a utility function $\overline{V}_i = Y_i$, where $i = 1,2$; and each is endowed with 10 units of $Y$ and $1\frac{1}{2}$ units of $X$. All other individuals $i = 3,4,\dots,I+2$ again have preferences given by (3) and are initially endowed with $10Y$'s and no $X$'s. (Note I am now assuming there are $I+2$ individuals in the economy, where $I+2 > 2$.)

Figure 4 again shows the market demand for securities, so in the absence of short selling, the firm faces a perfectly elastic demand for its shares and shareholders unanimously favor net market value maximization.

That is, they favor no production since production would entail a net loss of \$2: the marginal cost of an $X$, \$6, exceeds anyone's marginal valuation of an $X$, \$4. The competitive equilibrium with no short selling would involve autarky.

But if short selling were possible, then both initial owners would like the firm to produce, in spite of the consequent loss! To see this observe that if the firm produces $(-6,1)$, the input cost of individuals 1 and 2 is $3Y$ each; and a short sale of a share in the firm is a promise to deliver a unit of $X$ in period 2. So if individual 1 or 2 sells $1\frac{1}{2}$ shares short—the most he can sell short given his endowment of $X$'s—and also sells his initial holding of $\frac{1}{2}$ share in the firm, he can earn $8Y$ gross or $5Y$ net of input costs. He can thus consume $15Y$'s instead of only $10Y$'s (his consumption under autarky).

More specifically, in a competitive equilibrium with short selling, each individual except 1 and 2 consumes $4/I$ units of $X$ (costing $\$16/I$), and $10-16/I$ units of $Y$; each purchases $4/I$ shares yielding $1X$ per share and costing \$4 per share. Individuals 1 and 2 consume no $X$'s and $15Y$'s; each sells 2 shares ($1\frac{1}{2}$ short) earning \$8 or, net of input costs, \$5 each. Note the loss to individuals 1 and 2 from the firm's production (\$1 each), is more than compensated for by the gain from short selling (\$6 each). Also note that even with individuals 1 and 2 selling shares short, the firm faces a perfectly elastic demand for its shares. In particular, Figure 4 illustrates the security market equilibrium (now with short selling). Thus it is indeed the violation of Assumption 4* rather than a violation of the perfect competition assumption that leads to the failure of net market value maximization in this example.

The failure of market value maximization here is due to a peculiar type of externality: while a perfectly competitive firm's net market value *internalizes* all the benefits that accrue to *buyers* of the firm's shares (this follows from the no surplus characterization, (v), and is the intuitive reason underlying the validity of (i), it does not reflect *any* of the benefits that accrue to *short sellers*: they only deal with third parties and do not affect the firm's share price (in accordance with the perfect competition assumption). The firm only creates an *external* benefit for them by opening up the market for its shares.

A final note about the example. In this example both shareholders want the firm to produce at a loss. But in general there will be no unanimity without Assumption 4*, even under perfect competition. To illustrate this, change Example 5 so that only individual 1 has preferences $\overline{V}_i$ while individual 2 has preferences $V_i$. Then it is easy to check that if $I=3$, only individual 1 will want the firm to produce at a loss, individual 2 favoring no production and autarky.

## IV. Summary

I have shown that even if a firm's shares are unique, if a firm is a perfect competitor and there are no short sales of the firm's shares, then net market value maximization will be unanimously favored by its initial shareholders.

If the firm's shares are spanned by other existing securities, the no short sales restriction may be eliminated. But if they are not spanned and short selling of the firm's shares is possible, then there will in general be no unanimously favored objective for the firm (Example 5).

I have also shown that whether or not there is spanning or short selling, perfectly competitive conjectures can be correct in a general equilibrium environment in which firms have a positive net market value.

More sharply expressed, spanning is fundamentally irrelevant to the existence of perfect competition (Examples 2–5). And, in the absence of short selling, it is also fundamentally irrelevant to unanimity.

## APPENDIX: A BRIDGE TO THE FINANCE LITERATURE

In the finance literature, the perfect competition assumption is sometimes expressed that each individual conjectures that his implicit prices on state contingent income will be unaffected by any firm's change in its investment decision (for example, see David Baron's survey article). To form a bridge with this literature, it may be helpful to reexpress Assumption 9* in these terms, and to sketch an alternative proof of (i) which explicitly uses individuals' implicit prices. For simplicity, in the sequel I shall strengthen Assumption 4* to assert that there is no short selling of shares. The reader should easily be able to extend the argument to the case when there is short selling, but no one wants to go short in unique securities.

Following Baron's version of the "standard finance model," let $V_j(x_0)$ represent the value of firm $j$'s shares in a security market equilibrium corresponding to investment $x_0 \equiv (x_{10},\dots,x_{J0})$ by firms, where $x_{j0}$ represents firm $j$'s investment decision, $j=1\dots J$. Then, from the first-order conditions for an individual optimum, we know that for any individual $i$, $i=1\dots I$, and all firms $j=1\dots J$.

$$\text{(A1)} \quad V_j(x_0) \geqq \sum_s p_{is}(x_0) f_{js}(x_{j0}),$$

and a strict inequality implies $\alpha_{ij}(x_0)=0$, where $p_{is}(x_0)$ is $i$'s implicit price for income in state $s$, $s=1\dots S$, in the stock market equilibrium corresponding to investment $x_0$; $f_{js}$ is firm $j$'s production function in state $s$; and $\alpha_{ij}(x_0)$ is $i$'s *ex post* shareholdings in firm $j$, in the stock market equilibrium corresponding to $x_0$. Intuitively, (A1) just says that individual $i$ will buy shares in $j$ up to the point where his personal marginal valuation of shares in $j$ equals their market value. And if the latter exceeds the former even when $i$ holds none, then the no short selling constraint must be binding for $i$. (To derive (A1) simply follow Baron's argument to his equation (2), adding the complementary slackness condition resulting from our no short selling constraint.)

Similarly, if firm $j$ changed investment to $x_{j0}+\Delta x_{j0}$ while all other firms $j'$ remained at $x_{j'0}$, in the new security market equilibrium corresponding to $x_0+\Delta x_{j0}e_j$ (where $e_j$ represents the unit vector in $R^J$ with zeros except in the $j$th component), for all $i=1\dots I$,

$$\text{(A1}'\text{)} \quad V_j(x_0+\Delta x_{j0}e_j) \geqq \sum_s p_{is}(x_0+\Delta x_{j0}e_j) \times f_{js}(x_{j0}+\Delta x_{j0}),$$

and a strict inequality implies $\alpha_{ij}(x_0+\Delta x_{j0}e_j)=0$.

The perfect competition assumption of finance literature may now be expressed by saying $p_{is}(x_0+\Delta x_{j0}e_j)=p_{is}(x_0)$ for all $i$, $j$, and $\Delta x_{j0}$. If this assumption is *correct*, then (A1) plus (A1′) clearly imply that each individual $i$ conjectures for all $j$ and $\Delta x_{j0}$

$$\text{(A2)} \quad V_j(x_0+\Delta x_{j0}e_j) = \max_i \left\{ \sum_s p_{is}(x_0) f_{js}(x_{j0}+\Delta x_{j0}) \right\}.$$

Now (A2) can be used as a substitute for Assumption 9* since it essentially expresses the same idea. Note the intuition behind (A2): if the equals sign in (A2) were replaced by a greater than sign, no one would be willing to hold $j$'s shares, given (A1′) and the competitive assumption; if it were replaced by a less than sign, some individual's $p_{is}$'s would have to fall from $p_{is}(x_0)$ when $j$ changes to $x_{j0}+\Delta x_{j0}$ (recall (A1′)), contradicting the competitive assumption.

Now to prove (i) using (A2), first verify that given the competitive assumption there is no consumption effect, only a wealth effect, from any changes in investment. So, each initial shareholder $i$ in firm $j$ will favor $x_{j0}+\Delta x_{j0}$ to $x_{j0}$ if and only if the gain in share value exceeds its investment cost, that is, if and only if,

$$\text{(A3)} \quad V_j^i(x_0+\Delta x_{j0}e_j) \geqq V_j(x_0)+\Delta x_{j0},$$

where $V_j^i(x_0+\Delta x_{j0}e_j)$ represents $i$'s conjectures about $j$'s value in the security market equilibrium corresponding to investments $x_0+\Delta x_{j0}e_j$. (To verify (A3), follow the argument in the text, above; alternatively, for infinitesimal changes, follow Baron's argument to his equation 8.) Next, it is usual to complete the proof of unanimity by invoking spanning to show the $V_j^i$'s are the same for all individuals $i$; for example, see Baron's argument from his equation 8 to his equation 10. But we have already seen that if individuals' competitive conjecture are *correct*, then for all individuals $i$ $V_j^i(x_0+\Delta x_{j0}e_j)=V_j(x_0+\Delta x_{j0}e_j)$ as defined in (A2) above. That is, the line of proof using spanning is unnecessary.

## REFERENCES

**Baron, David P.,** "Investment Policy, Optimality, and the Mean-Variance Model," *Journal of Finance*, March 1979, *34*, 207–32.

**DeAngelo, Harry,** "Competition and Unanimity," *American Economic Review*, March 1981, *71*, 18–27.

**Diamond, Peter A.,** "The Role of the Stock Market in a General Equilibrium Model with Technological Uncertainty," *American Economic Review*, September 1967, *57*, 759–76.

**Drèze, Jacques H.,** "Investment under Private Ownership: Optimality, Equilibrium and Stability," in his *Allocation Under Uncertainty: Equilibrium and Optimality*, New York: Wiley, 1974, 129–166.

**Grossman, Sanford J. and Hart, Oliver D.,** "A Theory of Competitive Equilibrium in Stock Market Economies," *Econometrica*, March 1979, *47*, 293–329.

**Hart, O. D.,** (1979a) "Monopolistic Competition in a Large Economy with Differentiated Commodities," *Review of Economic Studies*, January 1979, *46*, 1–30.

______, (1979b) "On Shareholder Unanimity in Large Stock Market Economies," *Econometrica*, September 1979, *47*, 1057–83.

______, "Perfect Competition and Optimal Product Differentiation," *Journal of Economic Theory*, April 1980, *22*, 279–312.

**Kreps, D. M.,** "Three Essays on Capital Markets," mimeo., Institute for Mathematical Studies in the Social Sciences, November 1979.

**Makowski, L.,** (1980a) "A Characterization of Perfectly Competitive Economies with Production," *Journal of Economic Theory*, April 1980, *22*, 208–21.

______, (1980b) "Perfect Competition, the Profit Criterion, and the Organization of Economic Activity," *Journal of Economic Theory*, April 1980, *22*, 222–42.

______, (1980c) "No Surplus in Large Economies," Discussion Paper No. 49, Cambridge University, January 1980.

______, "Competitive Stock Markets," *Review of Economic Studies*, forthcoming.

**Ostroy, J. M.,** "The No-Surplus Condition as a Characterization of Perfectly Competitive Equilibrium," *Journal of Economic Theory*, April 1980, *22*, 183–207.

______, "A Reformulation of the Marginal Productivity Theory of Distribution," mimeo., University of California-Los Angeles, November 1981.

# Private Discrimination and Social Intervention in Competitive Labor Markets

*By* SHELLY J. LUNDBERG AND RICHARD STARTZ*

Do laws forbidding discrimination reduce allocative efficiency? A common thread in economists' discussions of equal opportunity laws has been a presumption that equal pay and/or quota constraints placed on firms act as transfer mechanisms which, as a rule, cause efficiency losses. The implicit model which commentators employ, however, is based on competitive markets with perfect information, so that it is unclear how discrimination could have arisen in the first place.[1] Our purpose in this paper is to consider the efficiency effects of equal opportunity type intervention in the context of a conventional model of discrimination; that is, one which does not depend upon differences in innate ability between groups of workers to produce differences in wages.

The economic analysis of labor market discrimination has produced two general types of models: "taste" discrimination and informational or "statistical" discrimination. Taste models, such as Gary Becker's prototype, produce wage differentials based on the preferences of majority employers, employees, and customers, but of a type which should not generally persist in competitive markets. Statistical models, on the other hand, demonstrate that treating two groups of workers differently may be the rational response of firms to uncertainty about an individual's productivity. In this case, persistent wage differentials may arise between workers with the same productivity who belong to different, identifiable groups, even in competitive markets.

We present a simple model of statistical discrimination and examine the effects of prohibiting group-specific treatment of workers on both net social product and the distribution of income. The agents are competitive firms who pay wages equal to the expected value of a worker's marginal product, conditional upon all information available to them, and income-maximizing workers who decide on the size of their human capital investments based on known wage schedules. Each worker is characterized by a level of innate ability, and by affiliation with one of two groups. Firms are able to assess the marginal product of members of one group more reliably than for the second group's members, and so offer different wage schedules. The main result is that the allocation achieved by rational agents in this labor market can be improved by prohibiting discrimination based on group membership.

## I. Imperfect Information and Discrimination

Consider two groups of workers, defined according to race, sex, or some other easily observable, exogenous, characteristic. Each group contains individuals with varying levels of ability or skill which determine their marginal products in any employment. Risk-neutral firms, though they know the density functions which describe the distribution of ability for each type of worker, cannot observe directly the marginal product of an individual $i$, $MP_i$. They do, however, observe a test score $T_i$ which is some function of the worker's marginal product and group membership, $I_i$.

The group index $I_i$ will enter this relationship if the testing procedure differs across groups; that is, is biased or less reliable for one or the other. Since all firms are equally

---

*University of Pennsylvania and National Bureau of Economic Research; and the Wharton School, University of Pennsylvania, respectively. We acknowledge helpful comments from Charles Brown and an anonymous referee.

[1] For example, Finis Welch, in the course of an illuminating discussion of affirmative action enforcement (1981), constructs a simple model in which equal pay for workers who differ in ability does indeed distort occupational choices and cause efficiency losses.

effective in assessing $MP_i$ via the test, a competitive equilibrium will involve paying each individual a wage, $w_i$, equal to the expected value of marginal product conditional on the test score and group membership; that is, $w_i = E(MP_i | T_i, I_i)$. The wage schedule $w(T_i)$ will generally be different for the two groups, though for each group the average wage will equal its average marginal product.[2]

Dennis Aigner and Glen Cain (1977) present a simple model of statistical discrimination, based on that of Edmund Phelps (1972), which illustrates the general characteristics of this approach. They assume that the exogenously given (normal) distributions of productivity in each of two groups of workers (black and white) are identical. The test scores which firms observe, however, are more reliable indicators of ability for whites than for blacks. Thus,

$$T_i^B = MP_i + \varepsilon_i^B; \qquad T_j^W = MP_j + \varepsilon_j^W,$$

where $\sigma^2_{\varepsilon^B} > \sigma^2_{\varepsilon^W}$.

It is straightforward to show that the equilibrium wage is a weighted average of mean productivity and the individual's test score, where the test score of a black worker is weighted more lightly than the test score of a white worker. The wage schedule $w^B(T_i^B)$ will have a smaller slope than the schedule $w^W(T_i^W)$, though mean wages will equal mean productivities, which are identical for the two groups. High-scoring blacks will thus be paid less than whites with the same test score; the reverse will be true for workers with low scores.

Does the equilibrium represented in Figure 1 constitute discriminatory treatment of black and white workers? A definition of economic discrimination is required to answer that question. The most literal meaning of "to discriminate" is "to differentiate," and differential treatment alone might be commonly accepted as evidence of discrimination. In this sense, the separate wage schedules faced by two groups of workers with identical distributions of productivity are discriminatory.

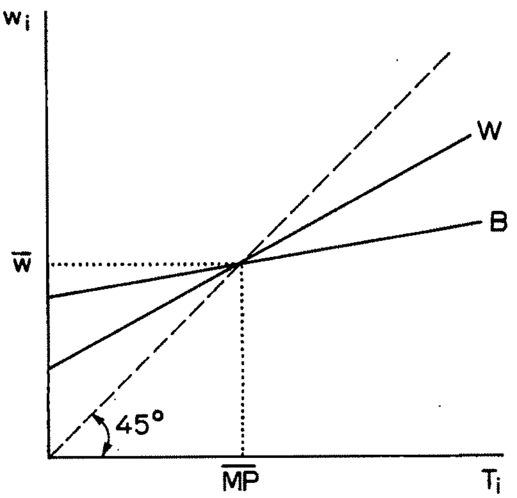


FIGURE 1

However, Aigner and Cain argue cogently that economists should not call this a discriminatory equilibrium, since groups with the same *average* productivity receive the same *average* compensation. Only a violation of this condition would be evidence of discrimination. In a perfectly competitive labor market, where firms pay each worker his expected marginal product according to an unbiased predictor, a nondiscriminatory equilibrium according to Aigner and Cain's definition is guaranteed.[3] Note that in a competitive market with perfect information, the absence of discrimination of the first sort is guaranteed as well. Only systematic differences in the quality of information available for the two groups of workers will cause rational employers to set different wage schedules.

---

[2] We assume that firms are able to assess average group productivity accurately, but not the productivity of individual workers. Wage differentials between equally skilled workers, however, may be allowed to erode slowly as each firm acquires information about individual productivities without changing the essential character of the model.

[3] Aigner and Cain describe a model which relies upon employer risk aversion to produce different mean wages for blacks and whites. This is discriminatory according to the criterion that groups with equal average productivity be paid unequal average compensation. However, each group is still compensated according to its contribution to the value of the firm, which now depends on risk as well as expected productivity.

Part of the disagreement over what constitutes discrimination can be traced to the fact that two different questions are being asked. The legal system typically wants to know whether a particular employer is rewarding all employees according to a single standard. Economists are concerned with whether the structure and operation of the labor market is such that workers are compensated efficiently (and perhaps equitably).

What implications do existing informational models of discrimination have for the efficiency of labor market equilibria? Since the wage schedules have no real effects on resource allocation (i.e., on labor supply or on the sorting of workers among jobs in which their productivities differ) the sole effect of discrimination is a redistribution of income among workers with the same level of ability. A crucial assumption is stated by Aigner and Cain, "Our focus is on labor market discrimination, which means we will generally assume that the worker's pre-labor market investments and endowments are given" (p. 177). Figure 1 thus represents a partial equilibrium model in which wages are competitively determined, but the actions and characteristics of workers are taken to be exogenous.

We propose to depart from this framework, and in doing so offer a generalization of the standard definition of economic discrimination.

DEFINITION: *Economic discrimination exists when groups with equal average initial endowments of productive ability do not receive equal average compensation in equilibrium.*

In this paper we recognize that human capital investment decisions will be affected by the presence of labor market discrimination and model this dependence explicitly. If wages are based on the results of an imperfect test, the returns to an investment in skills which cannot be directly observed are reduced. The result is a suboptimal level of human capital which will vary over groups if the quality of testing varies. We show in what follows that equilibrium allocations under this type of imperfect information are discriminatory and can be improved by such simple forms of labor market intervention as enforcing group-blind compensation rules.[4]

## II. Informationally Efficient Discriminatory Equilibria

We assume that workers have certain characteristics, both innate and acquired, which determine their productive abilities. These characteristics are distributed randomly in the population. Each worker knows his own characteristics exactly and invests in acquiring human capital to the point where the marginal cost of further investment just balances the increment to wages produced by the increased investment. It is important to note that human capital investments in our model do not consist merely of formal schooling, but of acquired abilities in the more general sense described by Kenneth Arrow: "Hence, the investments are not the usual types of education or experience, which are observable, but more subtle types of personal deprivation and deferment of gratification which lead to the habits of action and thought that favor good performance in skilled jobs..." (1973, p. 27).

Employers know the density function describing the distribution of characteristics through the population and observe a "test score" for each worker that provides information about the worker's marginal product,[5] but do not observe endowed or acquired human capital directly. The employer then offers a wage equal to the conditional expectation of the worker's marginal product. We set out below a simple stochastic model of worker characteristics and calculate the unique linear rational expectations equilibrium.

---

[4] The idea that anticipated labor market discrimination may affect education decisions has a long history in this literature. For example, it is a major theme in Welch (1967).

[5] This test score need not result from a literal "test," but is simply a summary measure of all information the employer is able to acquire during the hiring process and on the job. It may include, for example, years of formal schooling and years of work experience.

Each worker produces a marginal product $MP_i$ that depends on innate ability $a_i$ and acquired ability $X_i$. The desired level of $X_i$ can be purchased by each worker at increasing marginal cost, reflecting diminishing returns to time and money spent on training activities and increasing disutility of foregone leisure. Formulae for marginal product and the cost of acquired training are

$$(1) \qquad MP_i = a_i + bX_i;$$

$$(2) \qquad C(X_i) = .5cX_i^2, \quad C'(X_i) = cX_i.$$

In a full-information equilibrium, every worker purchases $b/c$ units of education at a cost $.5b^2/c$. The per capita net social product of education is

$$(3) \quad MP(X_i) - MP(0) - C(X_i) = .5b^2/c.$$

Employers do not observe true productivity, but rather a test score, $T_i$, for each worker. The test measures the worker's marginal product with a random error:

$$(4) \qquad T_i = MP_i + \varepsilon_i.$$

The worker characteristics $a_i$ and $\varepsilon_i$ are drawn from a bivariate normal distribution with known parameters $\bar{a}, \bar{\varepsilon}, \sigma_a^2, \sigma_\varepsilon^2$. We assume that $a_i$ and $\varepsilon_i$ are uncorrelated, though this assumption is not crucial. Workers know their own individual characteristics and maximize wages net of education costs. Employers are competitive and maximize profits by setting wages equal to the expected value of marginal product, conditioning the expectation on all available information. The parameters of the joint density function, as well as $b$ and $c$, are public knowledge.

There exists a unique linear rational expectations equilibrium for wages and human capital investments. In determining this equilibrium, workers look to the wage offer schedule to decide on the optimal level of human capital investment, and firms look to the joint distribution of test scores and marginal product to decide on the wage offer schedule. As a solution technique, we initially write optimal human capital investment as a linear function of worker characteristics with undetermined coefficients, as in (5) below. Properties of the equilibrium solution below allow us to fix unique values for the coefficients and thus completely characterize the equilibrium:

$$(5) \qquad X_i = \rho_0 + \rho_a a_i + \rho_\varepsilon \varepsilon_i.$$

The firm's problem is to establish a wage offer schedule as a function of test scores:

$$(6) \qquad w_i = E(MP_i | T_i) = E(T_i - \varepsilon_i | T_i)$$
$$= T_i - E(\varepsilon_i | T_i).$$

Since the test score is a linear function of normal random variables, the test score itself is normally distributed with mean $\bar{T} = b\rho_0 + (1 + b\rho_a)\bar{a} + (1 + b\rho_\varepsilon)\bar{\varepsilon}$ and variance $\sigma_T^2 = (1 + b\rho_a)^2\sigma_a^2 + (1 + b\rho_\varepsilon)^2\sigma_\varepsilon^2$. The test score $T_i$ and test error $\varepsilon_i$ have a bivariate normal distribution with correlation coefficient $(1 + b\rho_\varepsilon)\sigma_\varepsilon/\sigma_T$. The expectation of $\varepsilon_i$ conditional on $T_i$ follows immediately:

$$(7) \; E(\varepsilon_i | T_i) = \bar{\varepsilon} + \left[(1 + b\rho_\varepsilon)\sigma_\varepsilon^2/\sigma_T^2\right]\left[T_i - \bar{T}\right].$$

For convenience, we write the coefficient of the test score in (7) as $(1 - \beta)$. Substituting (7) into (6), we write the wage schedule offered by employers as

$$(8) \qquad w_i = \overline{MP} + \beta(T_i - \bar{T}).$$

Note that if $\bar{\varepsilon} = 0$, the individual wage is a simple weighted average of the group mean and individual test scores.

Each worker faces the wage schedule (8) with certainty and invests in human capital to the point where the marginal cost of acquiring more training equals the marginal increase in wages. An additional unit of $X$ increases the worker's marginal product and test score by $b$, so that wages rise by $\beta b$. The equilibrium level of acquired human capital is

$$(9) \qquad X_i = \beta b/c$$

for all workers.

Since the marginal cost of and marginal returns to each unit of $X$ are identical across workers, $X$ is nonstochastic in equilibrium—$\rho_a$ and $\rho_\varepsilon$ in (5) are identically zero. The equilibrium value of $\beta$ is

$$\beta = \sigma_a^2/\sigma_T^2 \tag{10}$$

so that $\beta$, which is also the ratio of marginal private to marginal social returns to acquired human capital, is between zero and one and depends only upon the relative sizes of the variances of innate ability and testing error. Since $MP_i = a_i + \beta b^2/c$ and $w_i = \overline{MP} + \beta[(a_i - \bar{a}) + (\varepsilon_i - \bar{\varepsilon})]$, it is easy to show that $\bar{w} = \overline{MP}$ and that $\sigma_w^2 = \beta\sigma_{MP}^2 = \beta\sigma_a^2$.

The net social product of education is

$$\begin{aligned} MP(X_i) - MP(0) - C(X_i) \\ = \beta(b^2/c)(1 - \beta/_2) \end{aligned} \tag{11}$$

so social welfare increases monotonically with $\beta$. Private markets result in an underinvestment in education.[6]

We now consider discriminatory equilibria. Suppose that workers are drawn from two subpopulations, the star group (*) and the dagger group (†). We assume the groups have identical mean innate characteristics $\bar{a}$ and $\bar{\varepsilon}$ and the same test variance $\sigma_T^2$. The only innate difference between the two groups is that the star group has relatively heterogenous innate ability and relatively homogeneous testing ability as compared to the dagger group.[7] Algebraically, $\sigma_a^2(*) > \sigma_a^2(\dagger)$ and $\sigma_\varepsilon^2(*) < \sigma_\varepsilon^2(\dagger)$. Using (10) we have $\beta^* > \beta^\dagger$.

Employers rationally discriminate between the star group and the dagger group by offering separate wage schedules. Workers in each group respond to their available opportunities and separate equilibria are calculated as above. The star group, whose test scores are more reliable indicators of productivity, becomes the high-wage/high-training group. Every member of the star group acquires $(\beta^* - \beta^\dagger)b/c$ more training than every member of the dagger group, since the marginal return to each unit of $X$ is higher. The average wage for each group, however, is equal to the group average marginal product.

Suppose the star group makes up $\alpha$ percent of the population and the dagger group the remaining $1 - \alpha$ percent. Total training is $(\alpha\beta^* + (1-\alpha)\beta^\dagger)b/c$ and the net social product of training is

$$\left[\alpha\beta^*(1 - \beta^*/2) + (1-\alpha)\beta^\dagger(1 - \beta^\dagger/2)\right]b^2/c. \tag{12}$$

Wage schedules for the star and dagger groups are reproduced in Figure 2. Does this situation constitute labor market discrimination? Regression tests of wages against test scores would reveal that the two groups are paid according to different schedules with star workers receiving larger raises for increased test scores, and higher average wages, than dagger workers. As Figure 2 shows, the average wage differential is more than can be explained by the difference in training using either the star or dagger schedule to calculate the wages due to increased training. In fact, for two workers with identical test scores, the star worker will generally receive the higher wage, the situation reversing only at very low test scores.

The legal system would almost certainly consider this equilibrium to be one of illegal discrimination, since dagger workers are generally paid less than star workers even after accounting for all observable individual characteristics. An economist, however, might disagree, since each worker is paid a wage equal to expected marginal product, and the average wage for each group is equal to that group's average marginal product. In our model, however, this is not an adequate test for discrimination, since productivities are

---

[6] Note that there is a first best policy the government could use to achieve social efficiency. It could order employers to use wage offer schedules with a coefficient of unity on test scores or subsidize wages at a rate of $(1 - \beta)$ per unit test score. However, the implementation of such a policy is not likely to be straightforward. We have assumed for simplicity that the test score and marginal product are measured in the same cardinal units, but this need not be true for the purposes of the model. To relax the restriction would require, for a first best solution, that the government know the relationship between test score and marginal product for each employer.

[7] Suppose, for example, that personnel managers are all members of the star group, and are more effective at assessing workers who are members of their own group.

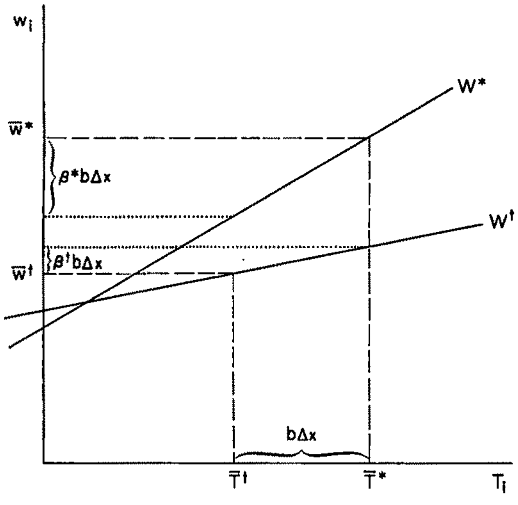


FIGURE 2

endogenous. Differences in group average productivities are a direct result of the incentives provided by employers and embodied in the wage schedule. Since different average wages are paid to groups whose average levels of "premarket" (in this case innate) abilities are the same, the labor market equilibrium depicted above is an example of economic discrimination as defined in Section I.

## III. Socially Preferable Nondiscriminatory Equilibrium

We demonstrate that a policy forbidding separate wage schedules for star workers and dagger workers results in an increase in allocative efficiency. Consider the consequences of the following policy restriction: employers may offer wages equal to the expectation of a worker's marginal product conditioned on his test score, but may not consider group membership.

We now derive the linear rational expectations equilibrium by arguments analogous to those in Section II. The key step is the derivation of $E(\varepsilon_i|T_i)$. A few intermediate calculations are required because the joint density function of $\varepsilon_i$ and $T_i$ is no longer bivariate normal. Let $f(\cdot)$ represent the density function of the mixture and $f^*(\cdot)$ and $f^\dagger(\cdot)$ represent the densities for the star and dagger groups. The density function of the mixture is

$$f(\varepsilon_i,T_i)=\alpha f^*(\varepsilon_i,T_i)+(1-\alpha)f^\dagger(\varepsilon_i,T_i). \tag{13}$$

Suppose, as we shall demonstrate, that there exists a linear rational expectations equilibrium. Let $\beta$ be the (as yet undetermined) coefficient of the test score in the wage offer equation. Every worker, star and dagger, will choose $\beta b/c$ units of education, so that $X_i$ is nonstochastic. The density functions $f^*(\cdot)$ and $f^\dagger(\cdot)$ are therefore bivariate normal as in Section II, except that star and dagger functions share a common mean test score which may differ from the mean test score of either group in the discriminatory equilibrium.

To find the conditional expectation, we need to find the conditional density function $f(\varepsilon_i|T_i)=f(\varepsilon_i,T_i)/f_T(T_i)$, where $f_T(T_i)$ is the marginal density function with respect to $T$. This marginal is a weighted sum of the star and dagger marginal densities, which are identical by construction, so

$$\begin{aligned} f_T(T_i) &= \int f(\varepsilon_i,T_i)\,d\varepsilon_i \\ &= \alpha\int f^*(\varepsilon_i,T_i)\,d\varepsilon_i+(1-\alpha)\int f^\dagger(\varepsilon_i,T_i)\,d\varepsilon_i \\ &= \alpha f_T^*(T_i)+(1-\alpha)f_T^\dagger(T_i) \\ &= f_T^*(T_i)=f_T^\dagger(T_i). \end{aligned} \tag{14}$$

The marginal distributions with respect to $T$ are $N(\bar{T},\sigma_T^2)$ for both the star and dagger group, and therefore for the mixture as well. The conditional density for the mixture is a weighted sum of the individual conditionals, given (13) and (14).

$$f(\varepsilon_i|T_i)=\alpha f^*(\varepsilon_i|T_i)+(1-\alpha)f^\dagger(\varepsilon_i|T_i). \tag{15}$$

Since expectation is a linear operator,

$$E(\varepsilon_i|T_i)=\alpha E^*(\varepsilon_i|T_i)+(1-\alpha)E^\dagger(\varepsilon_i|T_i). \tag{16}$$

Using the properties of bivariate normal distributions once again we have

$$(17)\quad E(\varepsilon_i|T_i)=\alpha\left[\bar{\varepsilon}+(1-\beta^*)[T_i-\bar{T}]\right]+(1-\alpha)\left[\bar{\varepsilon}+(1-\beta^\dagger)[T_i-\bar{T}]\right].$$

Notice that the conditional expectation of marginal product will be linear in the test score, even though the conditional distribution is nonnormal.

Conveniently, we can define

$$(18)\qquad \beta=\alpha\beta^*+(1-\alpha)\beta^\dagger,$$

and by setting the wage equal to expected marginal product, we can reproduce the linear wage schedule

$$(19)\qquad w_i=\overline{MP}+\beta(T_i-\bar{T}).$$

As a result of the policy restriction on wage schedules, all workers choose the same level of training. Average wages are the same for both groups and individuals with identical characteristics receive identical wages regardless of group membership. While the average marginal products of both groups are now equal, the variability in marginal product is greater for the star group. This implies that an individual employer faced with the equilibrium mixture still has an incentive to discriminate. The privately rational (but illegal) wage schedule pays each group the same on average, but, at high test scores, pays star workers more than dagger workers, and at low test scores, pays dagger workers more than star workers.

How does social welfare in the nondiscriminatory equilibrium compare with social welfare in the discriminatory equilibrium? By substituting (18) into (11) and comparing it with (12), we can see that total training is the same in both cases, but *net social product is higher in the restricted, nondiscriminatory equilibrium.* The improvement in social efficiency occurs because some high cost units of training have been shifted from star workers to dagger workers, for whom marginal training costs are lower. The ratio of nondiscriminatory to discriminatory costs is

$$(20)\qquad \frac{(\alpha\beta^*+(1-\alpha)\beta^\dagger)^2}{\alpha\beta^{*2}+(1-\alpha)\beta^{\dagger 2}}<1.$$

As an example, consider the case of maximum private discrimination, where $\beta^*=1$, $\beta^\dagger=0$, and the ratio of nondiscriminatory to discriminatory costs is $\alpha$.

## IV. Summary and Conclusions

The model constructed above provides a sparse representation of labor market discrimination. It produces differentials in average wages and rates of return to observable training between groups without appealing to differences in innate ability, risk aversion, or testing bias. We have shown that a competitive equilibrium under certain types of imperfect information can be improved by enforcing equal wage schedules for different groups of workers.

Our specific model can only give specific results. What more general lessons ought we draw about social policy toward discrimination? At a general level, the results of our paper are an example of the theory of the second best. In a first best world, economic agents would use all available information. In a second best world, there is no reason to assume that approaching the first best—using more information—is welfare improving. Since the problem of incomplete information is endemic in situations of discrimination, considerations of the second best are a general concomitant to policy questions in this area.

Specifically, our results arise because social marginal conditions diverge from private marginal conditions. Moving from the discriminatory to the nondiscriminatory equilibrium, we see that one group (the dagger workers) had its private incentives pushed closer to socially correct incentives, while the other group had its private incentives pushed even farther from the socially desirable level. The loss to society from the divergence between private and social incentives varies directly with the distance between the social and private incentives. The gain to society from discrimination, which reduces the small loss for the advantaged workers, is smaller than the loss to society from discrimination, which increases the large loss to the disadvantaged workers. While this result is not invariant with respect to specifications of cost and production functions, neither is it a peculiar case or the result of some special trick.

Our arguments have been intentionally and openly one-sided. We recognize the omission of social costs of antidiscrimination policy which might arise from production losses due to mismatches of workers and jobs, and from the costs to the government of maintaining any sort of social policy which must work against the private incentive structure. However, we believe we have demonstrated the need to be cautious in assessing the allocational consequences of equal opportunity type policies.

The appropriate social response to rational discrimination must be determined by analysis of specific problems. It is not an appropriate response, even for those of us who generally believe in the efficacy of private markets, to dismiss discrimination as something "the market will handle."

## REFERENCES

**Aigner, Dennis J. and Cain, Glen G.,** "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, January 1977, *30*, 175–87.

**Arrow, Kenneth J.,** "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973.

**Becker, Gary S.,** *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

**Phelps, Edmund S.,** "The Statistical Theory of Racism and Sexism," *American Economic Review*, September 1972, *62*, 659–61.

**Welch, Finis,** "Labor-Market Discrimination: An Interpretation of Income Differences in the Rural South," *Journal of Political Economy*, June 1967, *75*, 225–40.

______, "Affirmative Action and Its Enforcement," *American Economic Review Proceedings*, May 1981, *71*, 127–33.

# The Liquidity Premium and the Solidity Premium

*By* SUSAN WOODWARD*

The term structure of real interest rates is normally ascending. Given an ascending term structure, the implicit *forward* short-term rate of interest (for any specified future date) must be higher than the *current* short-term interest rate. Nevertheless, there is no evidence of any upward trend in real interest rates over time. It follows that implicit forward short-term rates must also be higher than the *actual* short-term rates that will be realized on average, in the future. This difference, between the implicit forward short-term rate and the mathematical expectation of the future short-term rate, is called the liquidity premium.

Thus two stylized facts: 1) an ascending term structure at any moment of time, and 2) historical stationarity of interest rates over time, imply a positive liquidity premium. Since historical stationarity might seem to need no special justification, explaining the liquidity premium reduces to explaining why the term structure is normally ascending. However, nonascending term structures are not rare. As for the historical stationarity of interest rates, empirically there seems to be no strong ground on which to postulate such a degree of uniformity. I therefore do not adopt the stylized facts as my point of departure. Instead, my intention is to identify the forces explaining the liquidity premium as such, independent of the shape of the term structure and the level of interest rates. Doing so will illuminate not only the determination of the liquidity premium but of the term structure as well.

John Hicks (1946) and J. M. Keynes (1930) argued that it is risk aversion which causes the forward rate to be greater than the expected future rate. Possible future variations in interest rates will affect the values of long-term bonds more than short-term bonds. Consequently, they argued, a yield premium is needed to induce risk-averse investors to hold the "less liquid" (more risky) longer-term bonds.

The Keynes-Hicks view has been criticized, and properly so, for its overemphasis on *capital-value risk* as opposed to *income risk*. Someone who wants only to assure himself a more-distant future flow of income could simply make a long-term investment commitment, and then be entirely unconcerned about possible interest rate variations and consequent fluctuations in capital values at intermediate dates. For such an individual, a yield premium might be required to induce him to hold a shorter-term instrument.

Such a consideration led Franco Modigliani and Richard Sutch (1966) to the notion of the "preferred habitat." They argued that because of personal variations in individual motives to save or to consume at different dates (due, for example, to life cycle considerations), different investors would typically be concerned about consumption risk at different dates. Thus, nothing of a high order of generality could be said about the sign of the liquidity premium.

While the Modigliani-Sutch criticism of Keynes and Hicks is well founded, their terminology may incorrectly suggest that each investor is typically concerned only with income risk at some single "habitat" date, or even that an individual's temporal consumption plans are exogenous data rather than endogenous variables. In the models to be presented here, individuals are concerned about income risks at every date (up to their planning horizons), and their temporal consumption plans are choice variables that depend upon relative prices and interest rates.

Once we acknowledge that individuals are risk averse with respect to income at *all* dates in their lifetimes, we must also accept that there is generally no single measure of the riskiness of an asset. For example, in comparing two pure discount, default-free bonds, one maturing at an earlier date and the other at a later date, the risk-configuration that first comes to mind (recognizing throughout that each bond yields a riskless return as of its own maturity date) is: (a) the longer-term bond is risky as of the earlier date, since its liquidation value will vary with intervening interest rate fluctuations, while (b) the shorter-term bond is risky as of the later date, since the investor's terminal return will vary with the rollover interest rate. The question "which maturity strategy is more risky?" has no single unambiguous answer.

---

*University of California-Los Angeles. Many thanks to Jack Hirshleifer, who carefully read and criticized every draft.

Suppose, on the other hand, that the terminal value of the short-term bond rolled over covaries negatively with the endowed consumption risk at the later date; that is, short-term bonds provide insurance with respect to the later date. Then buying the short-term bond and rolling over is even *less* risky with respect to the later date than holding a riskless bond maturing at that date. Here we have a configuration where, at the earlier date, the long-term bond is risky and the short-term bond is riskless, while at the later date, the long-term bond is merely riskless but the short-term strategy provides insurance. The short-term strategy is the less-risky strategy for obtaining consumption at *both* the earlier date and the later date! Hence in equilibrium, the short-term strategy must have a lower rate of return, on average, over both the near and distant future, than the long-term strategy. This means, of course, that the liquidity premium is positive.

In what follows, it will be shown that the liquidity premium will be positive if and only if the second risk-configuration described above prevails. I will also show there is a third configuration in which the long-term bonds are less risky, and therefore have on average the lower rate of return. A general equilibrium model will be employed to analyze how individuals' time-and-state distributed endowments, preferences (for consumption and also for risk bearing), beliefs (and the timing of information arrival affecting beliefs), and productive opportunities all contribute to the shaping of the term structure, and to the differential rates of return of short- and long-term bonds. I will first use a simplified world of pure exchange to highlight certain elements of the general theory of the liquidity premium, and will show, in particular, how the amount of and timing of anticipated information arrival affect the liquidity premium. I will then introduce intertemporal production. Here the interaction of production with information about the realized magnitude of representative endowments (for example, good or bad crops) affects beliefs about consumption in a systematic way, and plays a critical role in determining the liquidity premium.

## I. Definitions

We can cover all the essentials of the problem by dealing with a simple three-date model (dates 0, 1, and 2 years from the present). The current short-term real interest rate ${}_0r_1$ is defined in

$$(1)\qquad {}_0P_1/{}_0P_0 \equiv 1/(1+{}_0r_1).$$

Here ${}_0P_1$ is the price quoted today (date 0) of a unit claim to real income to be received next year (at date 1), and ${}_0P_0$ is the price today of a unit of real income today; I will ordinarily take current income of any date as the numeraire commodity for prices quoted at that date, so that ${}_0P_0 \equiv 1$. The interest rate denoted ${}_0r_1$ is then the rate quoted at date 0, today, for discounting one-year future claims into their current (present value) equivalent.

Analogously, the current long-term (2-year) interest rate ${}_0R_2$ is defined in

$$(2)\qquad {}_0P_2/{}_0P_0 \equiv 1/(1+{}_0R_2)^2.$$

Here ${}_0P_2$ is the price quoted today of a claim to income to be received at date 2, and ${}_0R_2$ is the corresponding long-term interest rate for discounting such claims into their present value equivalent. But the price ratio on the left-hand side of equation (2) can also be written in another way that serves to define the forward short-term rate ${}_0r_2$:

$$(3)\qquad {}_0P_2/{}_0P_0 \equiv (1+{}_0R_2)^{-2} \equiv [(1+{}_0r_1)(1+{}_0r_2)]^{-1}.$$

The notation ${}_0r_2$ indicates that this interest rate, though short term in that it discounts income claims from date 2 back only to the previous year (date 1), is a rate quoted (or else implicit in other quoted prices and interest rates) at date 0.

As the economy moves historically through time, at date 1 an actual short-term rate ${}_1r_2$ will come into existence for discounting date 2 claims. This is the future short-term rate, defined in terms of the prices and interest rate quoted at date 1:

$$(4) \qquad {}_1P_2/{}_1P_1 \equiv 1/(1+{}_1r_2).$$

(Here, since date 1 income becomes the numeraire for all claims quoted at date 1, it must be that ${}_1P_1 \equiv 1$.) Viewed from date 0, however, this actual future short-term rate ${}_1r_2$ will be uncertain. This brings us, to the formal definition of the liquidity premium $L$:

$$(5) \qquad L \equiv {}_0r_2 - E({}_1\tilde{r}_2).$$

The liquidity premium is the excess, on average, of the known forward short-term rate of interest over the unknown future short-term rate.

The simple algebra of the "discount" concept is completely analogous. The current short-term real discount ${}_0d_1$ is defined in

$$(6) \qquad {}_0P_1/{}_0P_0 \equiv 1+{}_0d_1.$$

Interest and discount are thus related as

$$(7) \qquad 1+{}_0d_1 \equiv 1/(1+{}_0r_1).$$

Evidently, positive interest will correspond to negative discount.

The forward discount is defined in

$$(8) \qquad {}_0P_2/{}_0P_1 \equiv 1+{}_0d_2 \equiv 1/(1+{}_0r_2),$$

and the actual future discount is defined in

$$(9) \qquad {}_1P_2/{}_1P_1 \equiv 1+{}_1d_2 \equiv 1/(1+{}_1r_2).$$

This brings us to the formal definition of what will be our second risk measure, the solidity premium, $S$ (J. Huston McCulloch, 1973 and Martin Bailey, 1964). The value $S$ is the difference between the forward discount and the expected future discount. Just as $L$ measures the riskiness of the short-term investment strategy as of the later date, $S$ measures the riskiness of the long-term strategy at the earlier date:

$$(10) \qquad S \equiv {}_0d_2 - E({}_1\tilde{d}_2).$$

Of course, $S$ can also be expressed in terms of interest rates, or in terms of price ratios:

$$(11) \qquad S \equiv 1/(1+{}_0r_2) - E(1/(1+{}_1\tilde{r}_2))$$
$$\equiv {}_0P_2/{}_0P_1 - E({}_1\tilde{P}_2/{}_1\tilde{P}_1).$$

A similar easy development from (5) leads to an expression for $L$ in terms of price ratios:

$$(12) \qquad L \equiv {}_0P_1/{}_0P_2 - E({}_1\tilde{P}_1/{}_1\tilde{P}_2).$$

Although there is a kind of inverse relationship, it will be evident that $S$ is neither the reciprocal nor the negative of $L$.

The intuitive interpretation that makes these formalisms useful can be developed as follows. Think of the liquidity premium $L$ as measuring the difference, as of the farther-future date (date 2), between the sure return (principal plus interest) on a long-term bond and the mean return on a short-term bond rolled over at the uncertain future short-term rate. For, it follows by manipulations that

$$(13) \qquad (1+{}_0r_1)L \equiv (1+{}_0R_2)^2$$
$$-(1+{}_0r_1)E(1+{}_1\tilde{r}_2).$$

As for $S$, think of it as measuring the difference, as of the nearer-future date (date 1), between the sure return on a short-term bond and the mean return on a long-term bond liquidated at the uncertain future discount. For

$$(14) \qquad S(1+{}_0R_2)^2$$
$$\equiv \frac{(1+{}_0R_2)^2}{1+{}_0r_2} - E\left(\frac{(1+{}_0R_2)^2}{1+{}_1\tilde{r}_2}\right)$$
$$\equiv (1+{}_0r_1)-(1+{}_0R_2)^2E(1+{}_1\tilde{d}_2).$$

Thus, $L > 0$ implies an excess yield, on average, of long-term over short-term instruments compared at date 2. On the other hand, $S > 0$ indicates an excess yield, on average, of short-term over long-term instruments compared at date 1. Since the dates at which the comparisons are made differ, $L$ and $S$ need not have opposite signs.

## II. Risk Premiums and the Return to Long- and Short-Term Bond Strategies

This section analyzes the forces underlying the liquidity premium and solidity premium, using an explicit contingent-claim model of income uncertainty at near-future (date 1) and far-future (date 2) dates. The present date is assumed free of uncertainty; at date 0 each individual is supposed to have a specific known endowment $\bar{c}_0$ of the real income commodity ("corn"). But, at date 1, his endowment will be the uncertain consumption vector $(\bar{c}_{11},\dots,\bar{c}_{1E};\ \pi_{11},\dots,\pi_{1E})$, where $e = 1,\dots,E$ indicates the state of the world at the earlier date and $\pi_{1e}$ represents the associated probability (assessed at date 0). Similarly, at date 2, the endowment will be $(\bar{c}_{21},\dots,\bar{c}_{2S};\ \pi_{21},\dots,\pi_{2S})$, where $s = 1,\dots,S$ is the index for states of the world at the later date and $\pi_{2s}$ is the associated probability in terms of beliefs at date 0. In this section, pure exchange is assumed: there are no productive opportunities (for example, storage) for physically transforming income of one date into income of any other date.

Suppose at date 0 there are complete markets for claims to consumption contingent upon states of the world, at any date.[1] Then the current price of a claim to income at date 1 contingent upon state $e$ can be denoted ${}_0P_{1e}$, and similarly ${}_0P_{2s}$ is the current price of a claim to income at date 2 contingent upon state $s$. Certainty claims to income as of any given date can be purchased by buying a full complement of the corresponding contingent claims. Thus,

$$(15)\qquad {}_0P_1 \equiv \sum_e {}_0P_{1e};\ {}_0P_2 \equiv \sum_s {}_0P_{2s}.$$

Once the passage of time reveals the state of the world at date 1, individuals will in general revise their beliefs about the probabilities of the date 2 states (i.e., the advent of state $e$ is not only an income event, but also generally an information event). The revised probabilities can be denoted $\pi_{s\cdot e}$. Then the price of a certainty claim to date 2 consumption, quoted at date 1 after state $e$ has obtained, can be expressed as

$$(16)\qquad {}_{1e}P_2 \equiv \sum_s {}_{1e}P_{2s\cdot e}.$$

The future short-term interest rate ${}_{1e}r_2$, contingent upon state $e$ having been realized at date 1, can be defined:

$$(17)\qquad {}_{1e}P_2/{}_{1e}P_1 \equiv 1/(1+{}_{1e}r_2).$$

As usual, the denominator on the left-hand side would be unity, since it is the price of the numeraire commodity current at date 1 after state $e$ obtains.

The expected future rate of interest, in terms of probability beliefs at date 0, can then be written

$$(18)\quad E(1+{}_1r_2) \equiv E\left(\frac{{}_1P_1}{{}_1P_2}\right) \equiv \sum_e \pi_e \frac{{}_{1e}P_1}{\sum_s {}_{1e}P_{2\cdot e}}.$$

Recalling equation (12), the liquidity premium $L$ can be expressed in terms of the underlying bundles of contingent claims as

$$(19)\quad L \equiv {}_0P_1/{}_0P_2 - E({}_1P_1/{}_1P_2)$$
$$\equiv \frac{\sum_e {}_0P_{1e}}{\sum_s {}_0P_{2s}} - \sum_e \pi_e \frac{{}_{1e}P_1}{\sum_s {}_{1e}P_{2s\cdot e}}.$$

---

[1] Specifically, I am assuming date 0 markets in $E$ distinct date 1 claims and, in $S$ distinct date 2 claims, $E+S$ markets. While more extensive regimes could be defined, for example, trading in all $E\cdot S$ claims contingent upon *both* a particular state at date 1 and a particular state at date 2, this assumption suffices for achieving preferred consumption vectors, that is, no Pareto-preferred improvements are made available by opening more markets. (After the realization of a particular state at date 1, retrading of the $S$ date 2 claims will in general take place.)

Before proceeding further, something more about the forces underlying the determination of prices should be said. First of all, I will assume away any differences of beliefs in the economy: everyone agrees as to the probabilities $\pi_e$, $\pi_s$, and $\pi_{s \cdot e}$. Let every individual make choices under uncertainty so as to maximize expected utility $U = E(V)$, where $V(c_0, c_1, c_2)$ is his "cardinal," risk-averse, preference-scaling function for dated consumption income vectors. Individuals are also assumed to be identical in endowments. Taken together, the assumptions of representative tastes and endowments for all individuals guarantee that the prices emerging in the model are "sustaining" prices—prices at which each agent merely holds his endowment and no trade takes place. In addition to the standard postulate of *state* independence of the utility function, I will also be making the simplifying assumption of *time* independence: that the $V$ function is separable in the variables $c_0$, $c_1$, and $c_2$. While this limits the generality of the results, nonseparable tastes (i.e., allowing for possible intertemporal preference complementarities) would be a second-order effect that can only be accommodated by rather burdensome notation.

If the social endowments of income are positive in every state, and if everyone assigns infinite negative utility to zero consumption at any date, an interior solution is guaranteed in which each individual holds positive amounts of every contingent claim $c_{1e}$ and $c_{2s}$. Then in equilibrium at date 0, each individual's expected marginal utilities will be proportional to the prices:

$$(20) \qquad \frac{v_0}{{}_0P_0} = \frac{\pi_e v_{2e}}{{}_0P_{1e}} = \frac{\pi_s v_{2s}}{{}_0P_{2s}},$$

where $v_0 \equiv \partial U/\partial c_0$, $\pi_{1e} v_{1e} \equiv \partial U/\partial c_{1e}$, and $\pi_{2s} v_{2s} \equiv \partial U/\partial c_{2s}$.[2] This equation reveals, therefore, the relation of prices to preferences and endowments (which together determine the marginal utilities) and to probability beliefs.

Substituting from (20) into (19) leads to

$$(21) \quad L = \frac{\sum_e \pi_e v_{1e}}{\sum_s \pi_s v_{2s}} - \sum_e \pi_e \frac{v_{1e}}{\sum_s \pi_{s \cdot e} v_{2s}}.$$

Or, in a more condensed notation,

$$(22) \qquad L = \frac{E(v_1)}{E(v_2)} - E\frac{v_1}{E_e(v_2)}.$$

Here $E_e$ indicates the expectation (of date 2 marginal utility) conditional upon state $e$ at date 1 (i.e., calculated in terms of the $\pi_{s \cdot e}$ probability beliefs). Expectations symbolized simply as $E$ are taken with respect to beliefs at date 0. Of course,

$$(23) \quad E(E_e(v_2)) = \sum_e \pi_e E_e(v_2) = E(v_2).$$

In parallel, a similar expression for $S$ is developed. Starting from equation (11), we obtain in a similar fashion the analog of equation (22):

$$(24) \qquad S = \frac{E(v_2)}{E(v_1)} - E\frac{E_e(v_2)}{v_1}.$$

Certain simple relations must hold between the signs of $L$ and $S$, following from Jensen's Inequality.[3] In particular, $L$ and $S$ may be both negative, but may not be both positive. Also $L > 0$ $(S > 0)$ implies $S \leqslant 0$ $(L \leqslant 0)$, $L = 0$ $(S = 0)$ implies $S \leqslant 0$ $(L \leqslant 0)$, but $L < 0$ $(S < 0)$ implies nothing about the sign of $S(L)$. Furthermore, only if both $v_1$ and $E_e(v_2)$ are nonrandom can both $L$ and $S$ be zero.[4]

---

[2] Equation (20) follows directly from maximization of $U = E(V)$ subject to the following budget constraint at date 0 (where the overbars indicate endowed quantities):

$$ {}_0P_0 c_0 + \sum_e ({}_0P_{1e} c_{1e}) + \sum_s ({}_0P_{2s} c_{2s}) = {}_0P_0 \bar{c}_0 + \sum_e ({}_0P_{1e} \bar{c}_{1e}) + \sum_s ({}_0P_{2s} \bar{c}_{2s}).$$

[3] Jensen's Inequality: $E(1/x) \geqslant 1/E(x)$; equality holds only for nonrandom $x$.

[4] If $L$ and $S$ were defined in terms of continuously compounded interest rates instead of discrete exchange ratios, then necessarily $L = -S$ so that $L$ and $S$ could not both be negative. But, as the case of $L$ and $S$ both negative has an interesting economic meaning, something is lost in going to the continuous-interest form.

Thus, the possible cases of interest are 1) $L>0$, $S<0$; 2) $L<0$, $S>0$; and 3) $L<0$, $S<0$. For the economic interpretation of these various cases, it is first necessary to show that $L$ and $S$ are explicit measures of risk. Note that $L$ and $S$ are not merely determined by the pattern of time-dated risks, but in fact can be rearranged to show that they can each be expressed as a covariance of an asset value with a dated marginal utility. Rearranging the expression for $L$ in equation (22):[5]

$$L = 1/E(E_e(v_2)) \cdot \times \left[E(v_1) - E\left(\frac{v_1}{E_e(v_2)}\right)E(E_e(v_2))\right]$$

$$= \frac{1}{E(v_2)}\left\{E\left(\frac{v_1}{E_e(v_2)}E_e(v_2)\right) - E\left(\frac{v_1}{E_e(v_2)}\right)E(E_e(v_2))\right\}.$$

which is

$$(25) \quad L = \frac{1}{E(v_2)} Cov\left(\frac{v_1}{E_e(v_2)}, E_e(v_2)\right);$$

and similarly,

$$(26) \quad S = \frac{1}{E(v_1)} Cov\left(\frac{E_e(v_2)}{v_1}, v_1\right).$$

The first term in the covariance in (25), $v_1/E_e(v_2)$, can be seen from equations (17), (18), and (20) to equal $1+{}_{1e}r_2$, that is, one plus the interest rate contingent on the occurrence of state $e$ at date 1. Similarly, its reciprocal, $E_e(v_2)/v_1$, the first term in the covariance in equation (26), is one plus the contingent discount.

From equations (25) and (26) we can derive three propositions:

PROPOSITION 1: $L>0$ *if and only if the covariance between future interest rates and date* 2 *conditional expected marginal utility is positive, or, roughly speaking, if and only if the covariance between interest rates and date* 2 *consumption is negative.*

This is the same thing as saying that $L>0$ when the value of a rolled-over short-term bond covaries against consumption at date 2, which makes the strategy of holding short-term bonds a form of insurance with respect to date 2. Recall that $L>0$ implies $S<0$; that is, that the covariance between the value of the long-term bond liquidated at date 1 and date 1 consumption is positive. The short-term bond strategy is thus the less-risky strategy for obtaining consumption at both the earlier and the later date, and hence has a lower rate of return over both the short term and the long term. This is scenario 1.

PROPOSITION 2: $S>0$ *if and only if the covariance between future discounts and date* 1 *marginal utility is positive, or, roughly speaking, if the covariance between future discounts and date* 1 *consumption is negative.*

Again, recall that $S>0$ implies $L<0$, or that the covariance between interest rates and date 2 consumption is positive. Thus, this scenario 2 occurs when it is the long-term bond that is the less-risky instrument for obtaining consumption at both dates. Hence the long-term bond has a lower rate of return than the series of short-term bonds, on average, over both the near and distant future.

PROPOSITION 3: $L<0$ *and* $S<0$ *if and only if the covariance between interest rates and date* 2 *conditional expected marginal util-*

---

Moreover, from an expository point of view, the interpretation of $L$ and $S$ in terms of covariances of bond values with consumption is much easier with the exchange-ratio definition. For commodities distinguished only by date, either the continuum or the discrete-period versions of time passage may be useful. But the discrete-period formalization lends itself more naturally to generalizations involving distinct commodities—for example, comparing the forward wheat-corn price ratio to the expectation of the future wheat-corn price ratio (see my 1980 paper)—since the trade of corn for wheat is necessarily discrete. For a discussion of the term structure using continuous compounding, see John Cox, J. E. Ingersoll, and Stephen Ross, (1979).

[5] The development that follows is due to McCulloch (1973).

*ity is negative and the covariance between discounts and date* 1 *marginal utility is also negative.*

In approximate terms, this case occurs when the long-bond liquidity covaries with date 1 consumption and the short-term bond rolled over covaries with date 2 consumption. This seemingly paradoxical case thus corresponds to the most easily imagined risk-configuration—that in which each bond is the least-risky strategy for obtaining consumption at its own maturity date (scenario 3).[6]

## III. The Stochastic Properties of Consumption and the Relative Rates of Return to Long- and Short-Term Bond Strategies

It has been shown that it is not the mere presence of risk that shapes the term premiums on bonds, but the pattern in which the resolution of time-dated risks is visited upon prices. Two aspects of the stochastic properties of consumption describe this pattern and determine the signs and relative magnitudes of $L$ and $S$: 1) the serial correlation of dated marginal utilities or—more approximately—the serial correlation of consumption over time, and 2) the relative magnitude of two coefficients of variation, the first of date 1 marginal utility and the second of date 2 marginal utility.

It will be shown in the next section how intertemporal productive transformations tend to induce positive serial correlation. In this section we shall see how arrival of information about date 2, either due to the mere realization of some state at date 1 or due to the arrival of some side-message, may affect the serial correlation either way. As for the coefficients of variation of the dated marginal utilities, these are reflections of the power of the date 1 information as pertaining to date 1 and to date 2. Since our knowledge about the date 1 state will be complete at date 1, the extent of the information gained is essentially measured by our date 0 uncertainty about the date 1 state—the coefficient of variation of $v_1$. But we may also at date 1 receive information about date 2. The coefficient of variation of $E_e(v_2)$ measures how much we expect (as of date 0) to revise (as of date 1) our initial beliefs about the date 2 state. Thus, both the relative *size* of the revisions (indicated by the coefficients of variation) and also the direction of *dependence* of the revisions (indicated by the serial correlation) affect $L$ and $S$.

Rearranging (25) and (26), and simplifying notation by letting $x \equiv v_1$ and $y \equiv E_e(v_2)$:

$$L = Cov(y, x/y)/E(y); \tag{27}$$

$$S = Cov(x, y/x)/E(x).$$

These expressions can be rearranged to obtain:[7]

$$L = E(x)\sigma_{1/y}\left[\frac{\sigma_y}{E(y)}\rho_{y,1/y} - \frac{\sigma_x}{E(x)}\rho_{x,1/y}\right]; \tag{28}$$

$$S = E(y)\sigma_{1/x}\left[\frac{\sigma_x}{E(x)}\rho_{x,1/x} - \frac{\sigma_y}{E(y)}\rho_{y,1/x}\right]. \tag{29}$$

Here $\sigma_x$ and $\sigma_y$ are the standard deviation of date 1 marginal utility and the standard deviation of conditional expected date 2 marginal utility, respectively; $\sigma_{1/x}$ and $\sigma_{1/y}$ are the standard deviations of their reciprocals; and $\rho$ is the coefficient of correlation.

---

[6] As for the $L > 0$, $S > 0$ case, this cannot occur. Note that $L > 0$ implies that the short-term strategy provides insurance at date 2, which means the high contingent interest rates must be associated with small expected consumption at date 2. Note that $S > 0$ implies that long-term bonds provide insurance at date 1, which means that high interest rates (discounts) must be associated with large consumption endowments at date 1. But, if the realized state at date 1 turns out to be rich, while date 2 is likely to be poor, the contingent interest rate would have to be low, not high.

[7] These expressions follow straightforwardly from the fact that

$$Cov(y, x/y) = E(x)\sigma_y\sigma_{1/y}\rho_{y,1/y} - E(y)\sigma_x\sigma_{1/y}\rho_{x,1/y}$$

and $$Cov(x, y/x) = E(y)\sigma_x\sigma_{1/x}\rho_{x,1/x} - E(x)\sigma_y\sigma_{1/x}\rho_{y,1/x}.$$

The sign of $\rho_{x,y}$ does not conclusively determine the sign of $\rho_{x,1/y}$, but the two will fail to take on opposite signs only in rather bizarre cases where one variable is highly skewed. For purposes of exposition, therefore, I shall carry on as though the signs of $\rho_{x,1/y}$ and $\rho_{y,1/x}$ are always the opposite of $\rho_{x,y}$. (As for $\rho_{y,1/y}$, the correlation of any variable with its own reciprocal is always negative.)

Subject to the imprecision of this approximation, we have now three additional propositions:

PROPOSITION 4: *Negative serial correlation in the marginal utility of consumption implies $L<0$ and $S<0$.*

More approximately, negative serial correlation in consumption insures that neither bond will command an obvious premium over the other—we have scenario 3 corresponding to Proposition 3. (Strictly speaking, as the equations indicate, we could have scenario 3 even with $\rho(x,y)>0$, but only if it is small and if one distribution or both are highly skewed.)

PROPOSITION 5: *Positive serial correlation in the marginal utility of consumption implies either $L>0$ or $S>0$.*

In approximate terms, positive serial correlation in consumption implies that either one bond or the other will command an obvious premium—that we will have either scenario 1 or 2, corresponding to Propositions 1 and 2.

PROPOSITION 6: *Given positive serial correlation in the marginal utility of consumption, $\sigma_x/E(x)>\sigma_y/E(y)$ implies $S>0$, and $\sigma_y/E(y)>\sigma_x/E(x)$ implies $L>0$.*

Proposition 6 tells us that as between scenarios 1 and 2, which of the two prevails depends upon the *magnitude* of the associated positive regression. Suppose occurrence of a good state at date 1 raises the probability of good states at date 2, but not by so much as to make the expected date 2 outcome better (in marginal utility terms) than the realized date 1 outcome. Thus, $v_{1e}<E_e(v_2)<E(v_2)$. (And, of course, there would then be a similar "regression toward the mean" effect in the event of a bad state at date 1.) The coefficient of variation for date 2 expected marginal utility, $\sigma_y/E(y)$, then tends to be small in comparison with the corresponding date 1 statistic, $\sigma_x/E(x)$. This leads, in equations (28) and (29), to the pattern $L>0$, $S<0$—scenario 1. Here the anticipated revision of beliefs regarding date 1 is, on average, more extensive than that regarding date 2. But if a good state at date 1 means, on average, an even better outcome at date 2 (and a poor date 1 state an even worse expected date 2 outcome), this "regression away from the mean" leads to scenario 2 with $L<0$, $S>0$. In this case the anticipated revision of beliefs regarding date 2 is, on average, more extensive, even though we get conclusive information about date 1 and only partial information about date 2.

An interest rate interpretation will also be helpful. If the magnitude of the positive regression is small (regression toward the mean), the contingent future short-term interest rate will be *low* in *rich* branches of the date-state consumption tree and *high* in *poor* branches of the tree. Here a good date 1 state implies a not quite as good date 2 state on average, and therefore low ${}_1r_2$—while a poor date 1 state implies a high ${}_1r_2$. Low rollover yields when you are rich, and high yields when you are poor means that the short-term maturity strategy has the superior risk characteristics, so the long-terms must unambiguously yield more, on average, over both long and short horizons ($L>0$, $S<0$). But, with regression away from the mean, the contingent future short-term rate is high when you are rich and low when you are poor; the short-term strategy has definitely inferior risk characteristics, and hence must always yield more, on average, again over both long and short horizons ($L<0$, $S>0$).

Finally, with negative $\rho_{x,y}$ a good (bad) date 1 state implies a band (good) date 2 on average, and thus a low (high) contingent interest rate. So the short-term strategy is risk enlarging at date 2 (implying $L<0$) while the long-term strategy is risk enlarging at date 1 (implying $S<0$). (Recall, however, that the combination $L < 0$, $S < 0$ may occur even if $\rho_{x,y}$ is positive so long as it is not very large.)

Note that the aspect of date 2 consumption that is important in determining the sign of the liquidity premium is not the variation in $c_2$ (date 2 consumption) or in $v_2$ (date 2 marginal utility), but the variation in $E_e(v_2)$—the date 1 revised expectation of date 2 marginal utility. The variation in $v_2$ constrains the variation in $E_e(v_2)$, of course; the variation in $E_e(v_2)$ can be as great as that of $v_2$ only if the information about date 2 arriving at date 1 is conclusive, and in all other cases must be smaller. The variation in $v_2$ does not matter, however, except to the extent that it is manifested in $E_e(v_2)$, and hence in interest rates, prior to date 2. Specifically, it follows from equations (25) and (26) that if no information is revealed at date 1 about date 2, (so that $E_e(v_2)$ is a constant), then $L = 0$ and $S < 0$, *even if there is much more consumption risk at date 2 than at date 1.*

Part of what I believe Hicks and Keynes had in mind as a source of the liquidity premium was that the far future was inherently more uncertain than the near future, a proposition few of us would dispute. But this is neither a necessary nor sufficient condition for $L > 0$. From equation (28) the needed conditions are: 1) that date 1 and date 2 consumption be positively correlated; and 2) that information arriving at date 1 is relatively more informative about date 1 consumption than date 2 consumption. The combined weight of these conditions is rather restrictive, so that in a world of pure exchange, there could be no clear presumption that the liquidity premium is positive. As we shall see, it is the possibility of intertemporal productive transformation, in association with information anticipated to be received at date 1, that more powerfully forces $L > 0$.

The analysis thus far has been only of bonds, that is, of certainty claims to income, either at date 1 or at date 2. But, more generally, for any asset representing a bundle of contingent claims, its date 0 price will vary so that its expected yield to any date will be determined by the covariance of its future value with consumption at that date. What we really have here is a multiperiod generalization of the single period Sharpe-Lintner-Mossin capital asset pricing model (CAPM)—in which the question "how risky is this asset?" has possibly as many answers as there are dates.[8] In the CAPM, the factor determining an asset's expected yield is its covariance with the market portfolio, where the value of the market portfolio is simply the realization of date 1 (end-of-period) consumption. In the model here, or rather its rearrangement to allow for risky bundles as well as bonds, the covariance of the asset values with the realization of consumption at each date determines the expected yield to that date.

## IV. The Influence of Intertemporal Production

The foregoing analysis of the determinants of $L$ and $S$ under pure exchange makes possible a condensed treatment of the consequences of introducing intertemporal productive transformations. Jack Hirshleifer (1972) argued that, when intertemporal production in the form of costless storage is possible, a unit claim to the inherently more flexible date 1 crop becomes (other things equal) more desirable than a corresponding claim upon the date 2 crop, tending to bring about an ascending term structure—and, he claimed, a positive liquidity premium. While it is true that the forward-only nature of intertemporal production always enriches the future at the expense of the present, and hence tends to make the term structure of interest rates ascending, I will show that it does not always imply $L > 0$.

I will discuss now two illuminating polar cases. In each, I distinguish between the endowment and the actual consumption at each date. It is assumed here that investment exhibits constant or diminishing returns, and that the marginal return to investment depends only on the amount invested, not on the endowments at either date. It is also assumed that production goes forward only;

---

[8] Since the first draft of this paper, an explicit multiperiod CAPM in terms of dated consumptions has been provided by Douglas Breeden (1979).

part of the current crop may be invested for consumption later, but there is no mechanism for transforming the future endowment into consumption today.

In the first case, which leads to $L > 0$, *the state realized at date 1 determines the endowment at that date, but does not modify the probabilities of the possible endowments at date 2.* It does, however, affect the amount individuals choose to invest at date 1. Here revised beliefs about the date 1 crop imply investment decisions that give additional information about date 2 consumption.

In the second polar case, which leads to $S > 0$, *there is only one possible crop at date 1, but different possible messages about what the endowment might be at date 2.* Here it is the message about date 2 that determines the optimal investment at date 1. Here revised beliefs about date 2 imply investment decisions that give us additional information about date 1 consumption. Below I provide a numerical example of each polar case (with production taking the form of simple constant returns storage).

For the first case (beliefs about the date 2 endowment unaffected by the outcome of date 1), it is evident that the better the date 1 outcome, the more will be invested at date 1. Then, given a diminishing-returns production function, we can easily establish that the covariance between $E_e(v_2)$ and contingent interest rates is positive. The greater the investment at date 1, the higher consumption at date 2, and so the lower is conditional expected marginal utility at the later date. Further, since investment exhibits diminishing returns, the higher the investment, the lower is the marginal rate of return and hence the contingent interest rate. So in those (relatively well-endowed) date 1 states where investment takes place, $Cov[E_e(v_2), {}_1r_2]$ is positive. For those date 1 outcomes so poor that no investment is undertaken, the value of $E_e(v_2)$ is invariant. The contingent interest rates in these corner-solution states will be higher than over the average of all states, and the invariant $E_e(v_2)$ will also be higher than its average over all states. Therefore, counting the corner-solution states as well, the covariance between contingent interest rate and $E_e(v_2)$ is positive. The conclusion is clear: the liquidity premium is positive. (A limiting case of $L = 0$ results if *all* states at date 1 are so poor that investment never takes place.) Of course, $L > 0$ implies $S < 0$.

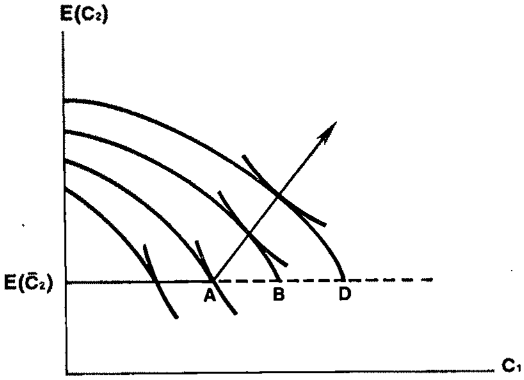


FIGURE 1

Figure 1 illustrates, for a representative individual situation, how a negative covariance between contingent interest rates and $E(c_2)$ arises. The latter is of course inversely related to $E_e(v_2)$, because the news changes only the mean of the $c_2$ distribution. As the date 1 endowment increases from very low levels, the expected date 2 endowment $E(\bar{c}_2)$ being constant throughout, no investment takes place until the critical point $A$ is reached. Beyond $A$, from endowment points like $B$ and $D$, the scale of investment increases. The interpretation, given diminishing returns to investment, is a negative association between the achieved $E(c_2)$ and the equilibrium contingent interest rate ${}_1r_2$.

In terms of equations (28) and (29), the first case reveals that *intertemporal production makes consumption positively serially correlated even if the endowments are independent.* A rich date 1 crop and hence a richer-than-average date 1 consumption now tend to be followed by enriched date 2 consumption, and a poor date 1 crop by a less-enriched date 2 consumption. Moreover, the low interest rates occur in the *rich* branches of the tree of outcomes, and the high ones in the poor branches.

This result, that intertemporal production implies $L > 0$, is very robust provided we adhere to the condition that beliefs about the date 2 endowment are unaffected. Notably, the positive liquidity premium does not depend upon the (perhaps more stylish than factual) stylized facts. Thus $L > 0$ obtains regardless of the distributions of $c_1$ and $c_2$ (and $c_0$), that is, whether or not the term structure is ascending (the first stylized fact). Further, $L > 0$ in no way depends upon a stationary trend in interest rates (the second of the stylized facts). The result will also hold for a paradigm with any number of dates, and for comparison of any maturities so long as all the dated endowments are statistically independent, and also for both linear and diminishing-returns production functions.[9]

As a numerical example of this polar case, consider an economy consisting of identical individuals whose tastes are described by $U = (c_0 c_1 c_2)$, and each of whom possesses investment opportunities in the form of costless storage between date 1 and date 2. (This simple example is constructed so that no storage would be chosen for date 0, even if it were possible.) The current (date 0) endowment is certain. There are two stages, $g$ (good crop) and $b$ (bad crop) at each future date. Prior beliefs are that each state at each date is equally probable, and the date 2 state is independent of the state realized at date 1. Endowments are then

| $t$: | 0 | 1 | 2 |
|---|---|---|---|
| | | 150 if $g$ | 150 if $g$ |
| | 90 | | |
| | | 75 if $b$ | 75 if $b$ |

Let us first consider what would happen in the absence of storage. In terms of equation (27), while there is variation in $x = v_1$ (the marginal utility of date 1 consumption), there is no variation in $y = E_e(v_2)$ (the conditional expected marginal utility of date 2 consumption). Hence we expect to observe $L = 0$, $S < 0$. Indeed, computation reveals ${}_0r_1 = .11$, ${}_0r_2 = 0$, ${}_0R_2 = .054$, $L = 0$, and $S = .125$.

For the first polar case, the information arriving at date 1 is conclusive regarding the crop at that date, but brings no news whatever about the date 2 crop. Such an information configuration implies for this example, that the utility-maximizing contingent consumption choices and corresponding storage decisions are

| $t$: | 0 | 1 | 2 |
|---|---|---|---|
| | | 125 if $g$ | 175 if $gg$<br>100 if $gb$ |
| | 90 | | |
| | | 75 if $b$ | 150 if $bg$<br>75 if $bb$ |

Evidently, there now is positive serial correlation between $x = v_1$ and $y = E_e(v_2)$ ($\rho_{xy} > 0$). The high interest rate is in the bad state, and the low interest rate in the good, implying $L > 0$. Computation based on the equality of ratios of marginal utilities and price ratios reveals ${}_0r_1 = .04$, ${}_0R_2 = .12$, and ${}_0r_2 = -.025$.

Now let us consider the second polar case:[10] here the crop at date 1 is nonstochastic, but information about date 2 will arrive at date 1 in the form of a side-message. This means that the date 1 investment decision depends only upon the news about date 2. The worse the news, the more will be set aside at date 1 (invested) for consumption at date 2. Note that uncertainty regarding date 1 *consumption* emerges even though there is certainty about the date 1 endowment. In effect, some of the risk at date 2 is shifted *back* to date 1.

Under the conditions of the second case, it is easy to see that the covariance between $v_1$ and $E_e(v_2)/v_1$ (one plus the contingent discount) will be positive, and hence by equation (26), the solidity premium must be positive. If the date 1 news about date 2 is very bad, investment is large and the contingent interest rate is very low. Correspondingly, since the date 1 endowment does not vary, date 1 consumption is very low. So, when investment occurs, lower interest rates are associated with lower date 1 consumption (implying high $v_1$). As for corner-solution states in which the date 2 news is so good that no investment is undertaken, in those states $v_1$ must be lower than average while interest rates are higher than average. Thus, overall, the covariance between date 1 marginal utility $v_1$ and the interest rate is negative. But ${}_1r_2$ is inversely associated with $E_e(v_2)/v_1$. So by equation (26) it must be that $S$ is positive, which implies $L<0$. (A limiting case of $S=0$ results if date 2 is always so well endowed that investment never takes place.)

---

[9] The liquidity premium is also positive with any linear production function. This can be shown by partitioning the covariance into storage (production) and no-storage (no-production) states, and showing that each partition is positive. The liquidity premium is also positive with a linear production function and independent or positively correlated endowments in infinite-time, two-state Markov models of the sort used by Robert Lucas (1978) and by Stephen LeRoy and Charles LaCivita (1980). It is easy to show that if there is merely persistence in the Markov transition probabilities, $L > 0$, and that with a linear production function added, plus either independence or persistence, $L$ is still greater than zero. However, if the transition probabilities already display persistence, the introduction of storage may not increase the size of $L$.

[10] I am indebted to John Riley for a classroom example which provided the basis for this idea.

As in the first case, the availability of intertemporal productive opportunities makes consumption serially correlated. But, here the high interest rates come in the rich branches of the tree of possible outcomes, and the low ones (along with large investment!) in the poor ones. Put another way, in the first case, higher investment (implying lower interest rates) takes place *from* richer states at date 1, while in our second case, higher investment takes place *toward* poorer states at date 2.

Figure 2 illustrates, for a representative individual, how a negative covariance between interest rates and $v_1$ arises. Along the horizontal axis $c_1$, date 1 consumption, is inversely related to $v_1$, the marginal utility of the date 1 outcome. On the vertical axis $\hat{c}_2$ represents the certainty equivalent of expected date 2 consumption. Since $c_2$ is still a random variable as of date 1, $\hat{c}_2$ is the analog of $c_1$, and is inversely related to $E_e(v_2)$. The date 1 endowment, $\bar{c}_1$, being constant throughout, as $\hat{c}_2$ increases investment undertaken at date 1 from endowment points like $E$ will decline until the critical point $F$ is reached. Beyond $F$ no investment at all takes place, but the slopes of the indifference curves along the vertical ray from $F$ are everincreasing. The interpretation is that, given diminishing returns to investment, there is a negative association between the consumed $c_1$ and the rate of interest.

As a numerical example of the second polar case, consider a situation where all of the crop uncertainty is at the later date, but the uncertainty is resolved at date 1, prior to the actual reaping of the crop. The representative individual has the same tastes as in the previous example, $U=(c_0c_1c_2)$. The preliminary endowment configuration is as follows, where the two states at date 2 are equally probable:

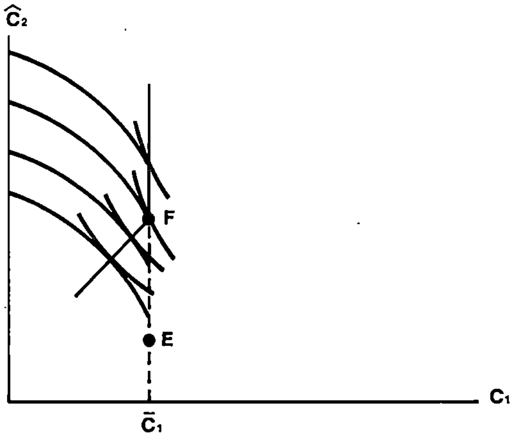


FIGURE 2

| $t$: | 0 | 1 | 2 |
|---|---|---|---|
| | 90 | 100 | 150 if $g$ |
| | | | 75 if $b$ |

First, if there is no storage option, there can be no variation in $x=v_1$, while the news at date 1 introduces variation in $y=E_e(v_2)$. Thus, $L<0$, but $S=0$. Numerically, ${}_0r_1=.11$, ${}_0r_2=0$, ${}_0R_2=.054$, $E({}_1r_2)=.125$, which gives $L=-.125$, $S=0$.

The information to arrive at date 1 conclusively determines the state at that date. The optimal choices become

| $t$: | 0 | 1 | 2 |
|---|---|---|---|
| | 90 | 100 | 150 if $g$ |
| | | 87.5 | 87.5 if $b$ |

Interest rates and term premiums are ${}_0r_1=.04$, ${}_0R_2=.11$, ${}_0r_2=.185$, ${}_1r_{2g}=.5$, ${}_1r_{2b}=0$, $E({}_1r_2)=.25$; $L=-.065$, $S=+.067$. Note that the higher of the future interest rates is realized in the "good" consumption stream, and the lower in the bad. This makes the long-term bond a less-risky asset from point of view of both future dates.

The second case illustrates that, from a sufficiently abstract point of view, even the availability of intertemporal production does not necessarily tend to make for a positive liquidity premium. Nevertheless, I argue that the general presumption in favor of a positive liquidity premium is not so wrong. The reason is that, apart from unusual situations, the information that individuals can anticipate receiving will typically do more to resolve uncertainty about near-future than about far-future events. The implication is that the first polar case above (in which nothing was learned about the date 2 endowment) is likely to be a closer approximation of reality than the second polar case (in which nothing was learned about the date 1 endowment). It is this first case, as shown, that led to a systematically positive liquidity premium.

## V. Summary and Conclusions

Why does anybody care about the sign of the liquidity premium? Because the term structure of interest rates alone does not reveal how risky it would be to provide for consumption at a given date by buying bonds that mature earlier (with the intention of rolling over), or by buying bonds that mature later (with the intention of liquidation). The liquidity premium $L$ measures the riskiness (with respect to consumption at a terminal date) of rolling over a series of short-term bonds. The solidity premium $S$ reveals the riskiness (as of an intermediate date) of the strategy of holding a longer-term bond and liquidating it.

The concern of Keynes and Hicks was for the investor who might, if he bought long-term bonds, have to liquidate them at an unfavorable price if he came upon a rainy day before they matured. In terms of the analysis here, this corresponds to $S < 0$—which means that the value of the long-term bond at an intermediate date (before its maturity) covaries *with* income at that date. When $S < 0$, the long-term bonds will be risk enlarging at intermediate dates, that is, they will indeed have their lowest values on rainy days and their highest values on sunny days. But $S < 0$ does not imply $L > 0$. This means that Keynes' and Hicks' investor could still be vulnerable even if he buys the short-term bond, as he may have to roll it over at an unfavorable rate. Only if $L > 0$ is the strategy of buying short-term bonds the less-risky strategy for both the near future and the far future.

Hicks' and Keynes' investor faces risk at two different future dates. Since consumption incomes at these two different dates are, in the Fisherian tradition, regarded as two distinct commodities, we need two measures of riskiness—one for each risky good—to fully describe the risk he faces.

For comparison of two bonds of different maturities, three basic scenarios can emerge:

1: Positive $L$ (which implies $S \leqslant 0$) occurs only if the short-term bond is the less-risky instrument with respect to both the earlier and the later future dates. This can happen only if the short-term bond provides insurance at the later date, implying (through Jensen's Inequality) that the long-term bond is risky with respect to the earlier date.

2: Positive $S$ (which implies $L \leqslant 0$) occurs only if the long-term bond is the less-risky instrument with respect to both dates; that is, it provides insurance with respect to the earlier date's endowed consumption risk.

3: Negative $L$ and negative $S$ occurs if both bonds are risky assets with respect to income at dates other than their own maturity. The expected yield on the longer-term bond at the shorter horizon is thus on average higher than the yield on the short-term bond; similarly the expected yield on the short-term bond rolled over must be higher than the yield on a long-term bond.

Which scenario emerges can be explained by ($i$) the serial correlation of consumption, and ($ii$) the relative sizes of the coefficients of variation of (a) marginal utility at the earlier date and (b) the conditional expectation (as of the earlier date) of marginal utility at the later date. The latter can be thought of as a measure of the relative informativeness of the earlier date's news regarding both the earlier date's and the later date's consumption.

If the serial correlation in consumption is zero or negative, both $L$ and $S$ must be negative. Then the least-risky instrument for any date's consumption is simply a riskless bond maturing at that date. But with positive serial correlation of consumption, one or the other of the two maturities becomes the less-risky instrument for obtaining income at both dates.

Should the correlation be positive, if the news arriving at the earlier date is on average more powerful regarding the earlier date than the later one (i.e., good news about the earlier date implies good, but not as good, news for the later, and similarly for bad news) the low interest rates will be in the richer branches of the tree of possible outcomes and the high interest rates in the poorer ones. Then the liquidity premium will be positive. Should the news regarding the later date be the more belief-revising (good news for the earlier date means great news for the later), the high interest rates come in the rich branches, the low in the poor, and the solidity premium is positive.

The positive liquidity premium observed by Keynes and Hicks is not a result of mere risk aversion. But, from the above, it can be shown to follow from risk aversion plus the positive serial correlation of consumption which arises from the forward-only nature of production, and the predilection of nature to give us more information about the near future than about the far future.

## REFERENCES

**Bailey, Martin J.,** "Discussion," *American Economic Review Proceedings*, May 1964, *54*, 554.

**Breeden, Douglas,** "An Intertemporal Asset Pricing Model With Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, September 1979, *7*, 265–96.

**Cox, John C., Ingersoll, Jr., J. E. and Ross, Stephen A.,** "A Reexamination of Traditional Hypotheses About the Term Structure of Interest Rates," unpublished paper, Stanford University, 1979.

**Hicks, John R.,** *Value and Capital*, 2nd ed., Oxford: Clarendon Press, 1946, 146–47.

**Hirshleifer, Jack,** "Liquidity, Uncertainty, and the Accumulation of Information," in C. F. Carter and L. J. Ford, eds., *Uncertainty and Expectations in Economics*, Oxford: Basil Blackwell, 1972, 136–47.

**Keynes, J. M.,** *A Treatise on Money*, Vol. II, New York: Harcourt, Brace & Co., 1930, 142–44.

**LeRoy, Stephen, and LaCivita, Charles J.,** "Risk-Aversion and the Dispersion of Asset Prices," *Journal of Business*, October 1981, *54*, 535–48.

**Lucas, Robert E.,** "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *46*, 1429–45.

**McCulloch, J. Huston,** "An Estimate of the Liquidity Premium," unpublished doctoral dissertation, University of Chicago, 1973.

**Modigliani, Franco, and Sutch, Richard,** "Innovations in Interest Rate Policy," *American Economic Review Proceedings*, May 1966, *56*, 178–97.

**Woodward, Susan,** "Properly Anticipated Prices Do Not, in General, Fluctuate Randomly," Working Paper, University of California-Los Angeles, 1980.

# Expectations and Reputations in Bargaining: An Experimental Study

*By* ALVIN E. ROTH AND FRANCOISE SCHOUMAKER*

One task of a theory of bargaining is to specify those factors which ultimately resolve the indeterminacy inherent in bargaining. Even if we expect bargainers to reach agreement in the region of individually rational, Pareto optimal contracts—which F. Y. Edgeworth (1881) called the contract curve—it is still necessary to analyze the factors contributing to the selection of a specific agreement.

Game-theoretic models rest upon the assumption that the outcome of bargaining among rational, fully informed agents is determined by the strategic possibilities available to the bargainers, and their preferences as represented by their expected utility functions. Indeed, games in which the players have this information are called games of *complete information*. A long-standing obstacle to testing the descriptive power of these models has been the requirement that the utility of the bargainers be known for each potential agreement. An experimental design permitting this problem to be circumvented in a laboratory setting was introduced in Roth and Michael Malouf (1979). In that experiment, the outcomes of bargaining under controlled conditions were observed to systematically deviate from the predictions of theory. Subsequent experiments, (Roth and Malouf, 1979; Roth, Malouf and J. Keith Murnighan, 1981; Roth and Murnighan, 1982) indicated that the outcome of bargaining is decisively influenced by factors other than the strategic possibilities and preferences of the bargainers, and isolated the cause of the observed deviations from the predictions of classical models, sufficiently so that a specific hypothesis could be proposed to account for them. This paper examines the hypothesis that those missing factors concern the subjective expectations of the bargainers about the behavior of their opponents. The results of the experiment reported here support the hypothesis and demonstrate that distinct, stable, self-fulfilling sets of expectations are compatible with a given bargaining situation as determined by the preferences and strategic possibilities of the bargainers.

## I. Review of Three Earlier Experiments

To test theories which depend on the expected utilities of the players, it is desirable to design experiments which allow the participants' utility functions to be determined. A class of games allowing this was introduced in Roth and Malouf. In that experiment, players bargained over the *probability* that they would receive some monetary prize, possibly a different prize for each player. Specifically, they bargained over how to distribute "lottery tickets" to determine the probability that each player would win his personal lottery (i.e., a player who received 40 percent of the lottery tickets would have a 40 percent chance of winning his monetary prize and a 60 percent chance of winning nothing). The rules of the game specified which distributions of lottery tickets were allowable. If no agreement was reached in the allotted time, each player received nothing. We call such games, in which each player has only two possible monetary payoffs, *binary lottery games*.

To interpret the outcomes of a binary lottery game in terms of each player's utility for money, recall that if each player's utility function is normalized so the utility for re-

---

*Mellon professor of economics, University of Pittsburg, and assistant professor of economics, University of Illinois, respectively. This work has been supported by NSF Grants SES 79-15356 and SOC 78-09928. We acknowledge the help of Michaël Barr and David Sides in setting up experimental procedures, and invaluable assistance from our colleagues, Lawrence M. DeBrock and J. Keith Murnighan, in analyzing our data.

ceiving his prize is 1, and the utility for receiving nothing is 0, then the player's utility for any lottery between these two alternatives is the probability of winning the lottery. The set of feasible utility payoffs in such a game equals the set of allowable divisions of lottery tickets. Thus binary lottery games can be used to experimentally test theories of bargaining which depend on the set of feasible utility payoffs. Note that the set of feasible utility payoffs does not depend on the prizes, so a binary lottery game in which the players know the allowable divisions of lottery tickets is a game of complete information, regardless of whether each player also knows the other's prize. The classical models of bargaining that follow the work of John Nash (1950) and depend only on the set of feasible utility payoffs to the players (see Roth, 1979) thus predict that the outcome of a binary lottery game will not depend on whether the players know their opponent's prize.

The experiment of Roth and Malouf was designed to test this hypothesis, among others. Participants played binary lottery games under either *full* or *partial* information. In the full information condition, each player knew his own potential prize and his opponent's potential prize. Under partial information, each player knew only his own prize. Contrary to the predictions of the classical models in the tradition of Nash (1950), the outcomes observed in the two information conditions exhibited dramatic differences: under partial information, outcomes tended to be very close to an equal division of the lottery tickets, while under full information, outcomes showed a pronounced shift toward equal expected payoffs.

Of course, other classical models describe games in greater detail. The *strategic* form of a game describes not only the feasible utility payoffs, but also the strategies available to the players. In the games described above, strategy choices concern the formulation of messages and proposals during negotiations. Since the strategies available to the players depend on the information they possess, we must consider whether the observed results can be accounted for by the different strategies available to the players in the two information conditions.

The experiment of Roth, Malouf, and Murnighan was designed to address this question and involved binary lottery games with prizes stated in terms of an intermediate commodity. Prizes were expressed in chips having monetary value, and each player played four games under either *high*, *intermediate*, or *low* information. In each condition, each player knew the number of chips in his potential prize and their value, but a player's information about his opponent's prize varied with the information condition. In the high information condition, each player knew the number of chips in his opponent's prize and their value. Under intermediate information, each player knew the number of chips in his opponent's prize, but not their value. Under low information, each player knew neither the number of chips in his opponent's prize, nor their value. In the latter two conditions, players were prevented from communicating the missing information about the prizes.

The experiment took advantage of two kinds of strategic equivalence relations. Binary lottery games with prizes expressed in both chips and money, played in the low information condition of this experiment, are strategically equivalent[1] to binary lottery games with the same monetary prizes expressed in money alone, played in the partial information condition of the previous experiment, because under the rules of the two information conditions, any legal message in one kind of game would be legal in the other. So the strategy sets are the same for both kinds of games, as are the utility functions and the underlying set of alternatives. Also, games expressed in both chips and money, played in the intermediate information condition of this experiment, are strategically equivalent to games expressed in money alone played in the full information condition of the previous experiment, so long as the monetary values of the two prizes in each money game are in the same proportion as the numbers of chips in the prizes in the corresponding chip game. This is because any legal message in one kind of game can be transformed into a legal message in the other kind of game by substituting references to chips for references to money (or vice versa) in any message concerning the value of the prizes.

---

[1] Two games are strategically equivalent if they can both be represented by the same game in strategic form. Thus, any theory of games which depends only on the strategic form of a game yields the same prediction for strategically equivalent games. (Compare Roth, Malouf, and Murnighan, 1981.)

If the observed difference between the partial and full information conditions of the previous experiment was due to the players' different strategy sets in the two conditions (the "strategic hypothesis"), then a similar difference should be observed between the low and intermediate information conditions of this experiment. The observed results did not support the strategic hypothesis. The low and high information conditions replicated the partial and full information conditions of the previous experiment, but the outcomes observed in the intermediate information condition did not differ significantly from those in the low information condition. In the intermediate information condition, the observed agreements tended to give both players equal probabilities, regardless of the size of their prize in chips. Thus, information about the artificial commodity, chips, did not affect the outcomes in the same way as did strategically equivalent information about money.

Both the above experiments revealed an effect of information which cannot be explained by existing models. The experiment of Roth and Murnighan was conducted to separate this effect into components resulting from the possession of specific information by specific individuals, and to assess the extent to which the observed behavior can be characterized as equilibrium behavior.[2]

In the two earlier experiments, either both bargainers knew their opponent's prize, or neither bargainer knew his opponent's prize. Also, it was always common knowledge whether the bargainers knew one another's prizes. Information is common knowledge in a game if it is known to all of the players, and if every player knows that all the players know, and that every player knows the others know that *he* knows, and so forth (compare David Lewis, 1969; Robert Aumann, 1976; and Paul Milgrom, 1981). Two bargainers can be thought of as having common knowledge about an event if the event occurs when both of them are present to see it, so that they also see each other seeing it, etc. In these experiments, a set of instructions provides common knowledge to the bargainers if it contains the information that both of them are receiving exactly the same instructions.

Each game of the third experiment was a binary lottery game in which one player had a \$20 prize and the other a \$5 prize. In each of the eight conditions of the experiment, each player knew at least his own prize. The experiment used a 4 (information)×2 (common knowledge) factorial design. The information conditions were: 1) *neither knows* his opponent's prize; 2) the *\$20 player knows* both prizes, but the \$5 player knows only his own prize; 3) the *\$5 player knows* both prizes, but the \$20 player knows only his own prize; and 4) *both players know* both prizes. The second factor made this information common knowledge for half the bargaining pairs, but not common knowledge for the other half.

The results of this experiment permitted three principal conclusions. First, the effect of information on the agreements reached is primarily a function of whether the player with the smaller prize knows both prizes. Second, whether this information is common knowledge influences the frequency of disagreement, with more disagreements occurring in the noncommon knowledge conditions. Third, in the noncommon knowledge conditions, the relationship among the outcomes showed virtually no departure from equilibrium.

Together, these three experiments permit us to speculate fairly specifically on the cause of the observed information effects. The first experiment demonstrated an effect of information about the prizes which could not be accounted for in terms of players' preferences over consequences (lotteries). The second experiment showed that this effect could not be accounted for by the set of available actions (strategies). The third experiment showed that the effect is consistent with rational (equilibrium) behavior. Thus, if we continue to hypothesize that the players are (approximately) Bayesian utility maximizers, it must be that the effect of information is due to a change in the players' subjective beliefs. For example, information about the prizes, and whether this information is common knowledge, may influence the players' subjective probabilities concerning what agreements are likely to be acceptable to their opponents.

---

[2] We refer here to the familiar *noncooperative* equilibrium of Nash (1951).

To see how the bargainers' expectations might influence the outcome, consider the following "thought experiment." A randomly selected individual plays some large number of games in which he bargains over how to divide a certain sum of money. Although he doesn't know it, all his opponents are confederates of the experimenter, and they all allow him to obtain, say, 80 percent of the available money. After he has gone through this experience, *you* have the opportunity of bargaining with him on your own behalf (i.e., not as a confederate). His past success is common knowledge. It will be difficult to bargain with him on an equal basis, since he expects (and has every reason to expect) to receive 80 percent of the available money, and since he expects (and has every reason to expect) that you will concede it to him. Suppose the rules of the game are that, after completing any negotiations, the players each separately write down their demands. They receive their demands if they are compatible, and otherwise receive nothing. If this is the only time you will be bargaining with him, the fact that this randomly selected individual now expects to get 80 percent will make it very risky for you to write down a demand of more than 20 percent.

To make precise how such subjective expectations can affect the decisions made by bargainers, consider a simple model of bargaining in a two-stage binary lottery game. In stage one, each individual $i$ makes a demand: he states the probability $p_i$ (of winning his prize) which he wants and offers $1-p_i$ to his opponent. In stage two, each bargainer chooses between repeating his demand or accepting his opponent's offer. An agreement occurs whenever the probabilities demanded add up to no more than 1. No messages can be exchanged.

If, in stage one, the two bargainers' demands add up to no more than 1, an agreement is reached at which each bargainer $i$ wins his prize with probability $p_i$. If the probabilities in stage one add up to more than 1, the outcome depends on the players' decisions in stage two. If each repeats his demand, a disagreement results and each will have a zero probability of winning his prize. If player $i$ repeats his demand $p_i$ and player $j$ accepts his opponent's offer, there is an agreement on $i$'s terms: $i$ will win his prize with probability $p_i$ and $j$ will win his prize with probability $(1-p_i)$. If both $i$ and $j$ accept the other's offer, $i$ will win his prize with probability $(1-p_j)$ and $j$ will win his prize with probability $(1-p_i)$.

Having both stated their demands in stage one, each bargainer must decide in stage two whether to repeat his demand or accept the other's offer. His expectations as to his opponent's behavior play a crucial role. The hypothesis which the experiment described in the next section is designed to test is that the expectations of the bargainers can be manipulated independently of the strategic possibilities and feasible outcomes of the bargaining situation. This is at odds with the traditional view, perhaps most explicitly stated by John Harsanyi (1977), who considers two-stage bargaining games of essentially this form, and argues that rational players' subjective probabilities are determined by the data of the game.

## II. The New Experiment

Agreements in the previous experiments tended to cluster around two divisions of the lottery tickets: one kind of agreement split the lottery tickets equally between the bargainers, the other gave the bargainers equal expected monetary payoffs. This experiment investigates whether, by manipulating the bargainers' expectations, one or the other of these two kinds of agreements can be obtained as a stable equilibrium.

Each player played 25 identical two-stage binary lottery games, of the kind discussed above. Although players were told that they bargained with another individual in each game, each individual in fact played against a programmed opponent (the computer) in the first 15 games, as in the thought experiment. Half the participants had a prize of $40 and half a prize of $10; both players knew both prizes, and players whose prize was $40 always bargained against players whose prize was $10 (each player had the same prize in all 25 games). Subjects were divided into three experimental conditions. The first was a 20–80 condition in which the computer was programmed to promote a 20–80 division of the lottery tickets, which yields equal expected payoffs. The second was a 50–50 condition in which subjects bargained with a computer programmed to promote the equal division of lottery tickets. The third condition was the control: subjects never bargained with the computer but always with other members of that group.

In the 20–80 condition, the agent whose prize was $40 bargained with a computer programmed to randomly select a first demand between 75 percent and 80 percent, and to repeat its demand in the second stage. The programmed opponent of the $10 player randomly selected a demand between 20 and 25 percent, and in stage two it accepted any offer giving it at least 20 percent of the lottery tickets. In the 50–50 condition, the programmed opponent of the $40 player randomly selected a first demand between 70 and 75 percent of the lottery tickets and in stage two accepted any offer giving it at least 50 percent. The $10 player bargained with a computer that randomly selected a first demand between 45 and 50 percent of the lottery tickets, and in stage two always repeated its demand.

In trials 16 to 25, subjects in each group bargained with other members of that group. Each game was played with a different anonymous opponent. Bargainers also received additional information, about their opponent's reputation as established in trials 11 through 15. They were told their opponent's first demand, whether he repeated it or accepted his opponent's offer, and which agreement, if any, was reached in each of trials 11 to 15, that is, in the final 5 games played against the computer, (when each player's behavior reflected his experience against a programmed opponent). In trials 16 to 25, every bargainer's experience against the programmed opponents was made common knowledge in this way, as in the thought experiment.[3]

Two elements influence the expectations of a pair of bargainers: their experiences and their reputations. Consider a typical pair in the 20–80 condition, at trial 16. The $40 player has never obtained an agreement in which he received more than 25 percent of the lottery tickets. He knows his opponent is aware of his reputation as established in games 11 to 15. Thus a $10 player, whose own experience has led him to expect that his opponent will capitulate to a demand for equal expected monetary payoffs, has his expectations reinforced when he confronts a $40 player whose reputation indicates that he has, in the past, given in to such demands. Similarly a $40 player, whose experience is that opponents are adamant about an equal division of expected payoffs, will have his expectations confirmed by a reputation indicating that his current opponent has previously behaved in this way.

The experiment is designed to distinguish between two competing hypotheses. The classical game-theoretic hypothesis, which states that the outcome of a game can be predicted from the set of feasible utility payoffs and strategic possibilities, implies that, in this experiment, the different experimental conditions should have no continuing effect. Specifically, if this is correct, we would expect to observe that, starting with trial 16, any differences between the two experimental conditions and the control condition would begin to disappear and the outcomes in the three conditions should converge over time, as continued play removes any transient effects due to the initial experience of players in the 20–80 and 50–50 conditions.

---

[3] Players were not informed, until after trial 15, that their reputations would be displayed. So reputation in this experiment serves only as an indicator of a player's past bargaining experience (in periods 11–15). This is in contrast with models in which the players know in advance that they are building a reputation. Robert Rosenthal (1979), and Rosenthal and Henry Landau (1979) consider repeated games with complete information in which players can strategically build and maintain reputations in the course of play, in order to influence future encounters. David Kreps and Robert Wilson (1982b), Milgrom and John Roberts (1982), and Kreps et al. (1982) consider games of incomplete information in which players may seek to play in early encounters in such a way as to build a reputation that will mislead future opponents about their true objective function. However, in the games played in this experiment, the fact that players are not aware that they are building a reputation in periods 11–15, and cannot subsequently alter the reputation established in those periods, removes any possibility that an incentive to alter his reputation can influence a player's bargaining behavior.

But if the expectations have a critical role in determining the outcome, as suggested by the earlier experiments, then we should expect to see divergent outcomes, established in the first 15 trials, persist in a stable fashion in each of the three conditions. We would expect the first condition's mean agreement to be near 20–80 and the second condition's to be near 50–50. The control condition's mean agreement should be somewhere between these two. This would be consistent with the hypothesis that the players' expectations were the uncontrolled factor accounting for the results of the previous experiments.

## III. Method

Each participant was seated at a visually isolated terminal of the PLATO computer system at the University of Illinois. Participants received all instructions and conducted all communication through the terminal. There were ten participants in each of the three conditions, which were conducted simultaneously. Subjects were undergraduates. Pretests were run with the same subject pool (but different subjects) to ensure that the instructions were easily understandable.

Background information including a brief review of probability theory was presented first. The rules of the game were then introduced. A demand $p_i$ was a number which was the sender's probability of winning his prize; $1-p_i$ was the probability offered to his opponent. As in the previous experiments, probabilities were presented in terms of the division of lottery tickets. The PLATO system computed the expected value of each demand and associated offer for both bargainers. After being made aware of these computations, a bargainer was given the option of cancelling his demand before its transmittal. After both demands were transmitted, stage two began. In stage two, each player had the choice of repeating his own demand or accepting his opponent's offer. When both decisions were made, the bargainers were informed of the outcome.[4] Participants were told that they bargained with a different individual in each game. In both stages, the bargainers were not informed of their opponent's decision until their own decision had been transmitted.

The bargainers in the 20–80 and 50–50 condition were paired with the appropriate programmed opponent for the first 15 trials. The instructions led them to believe they were bargaining with other individuals. The members of the control group were paired with other members of the control group. After trial 15, new instructions appeared on the screen and introduced the notion of reputation. Note again that as a bargainer was establishing his reputation, in trials 11 to 15, he did not know he was doing so. Trial 16 began with each bargainer's screen displaying both his own reputation and his opponent's. Each $40 player bargained twice with each $10 player. The value of the prizes was common knowledge.

After game 25, the payoffs were computed as described in the initial instructions: for each bargainer, one of games 1 to 25 was randomly selected; the corresponding lottery was then conducted with the specified prizes and the probabilities agreed upon in that randomly selected game. The players were directed to the monitor who paid them.

## IV. Results

Figure 1 shows the mean agreements, by trial, for each of the three conditions for the $40 player. The figure makes clear that the

---

[4] There was a time limit of two minutes for the first stage and one minute for the second stage.

TABLE 1—MEAN OUTCOMES: ALL NEGOTIATIONS INCLUDED; TRIALS 16–25

| | 20–80 | Control | 50–50 | Mean |
|---|---|---|---|---|
| \$40 Player | $22.4_c$ | $30.2_{bc}$ | $40.8_b$ | 31.1 |
| \$10 Player | $57.0_a$ | $39.0_b$ | $41.1_b$ | 45.7 |
| Mean | 39.7 | 34.6 | 40.9 | |

*Note:* Cells with common subscripts are *not* significantly different from one another at the .05 level using the Newman-Keuels procedure.

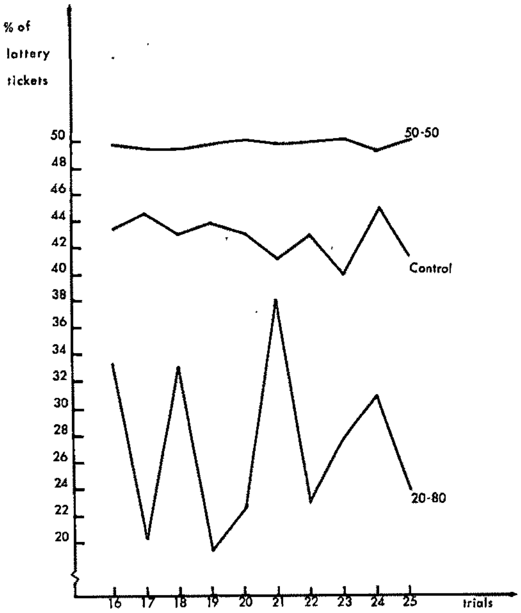


FIGURE 1. AVERAGE PERCENTAGE OF LOTTERY TICKETS OBTAINED BY \$40 PLAYER WHEN AGREEMENT WAS REACHED IN TRIALS 16–25

agreements reached in the three conditions differ markedly.

Preliminary analyses yielded no significant effects for trials for the total set of bargaining outcomes and for the set that excluded disagreements. Thus, the remaining analyses combined the data from the ten bargaining sessions. Analyses of variance for players (high prize vs. low prize) and condition (20–80, 50–50, and control) were conducted for the outcomes over all negotiations (including disagreements), for the outcomes excluding disagreements, and for the first and second offers made by the players. All the analyses yielded significant effects for condition and for the interaction between players and reputation ($F$-ratios in each case exceeded 16, with $p < .001$ in each case). The mean outcomes are shown in Tables 1 and 2, along with the results of *post hoc* tests on the significant interactions indicating that, in each case, outcomes in the 20–80 condition were significantly different from the 50–50 or control condition. The agreements that were reached in the control condition (with disagreements excluded) were between those of the other two conditions. The compelling nature of a 50–50 agreement in the 50–50 condition is highlighted by the extremely low variance in the 50–50 condition in contrast with the high variance in the 20–80 condition (see Table 3, which lists all final outcomes).

TABLE 2—MEAN AGREEMENTS: DISAGREEMENTS EXCLUDED; TRIALS 16–25

| | 20–80 | Control | 50–50 |
|---|---|---|---|
| \$40 Player | $27.3_e$ | $43.2_d$ | $49.7_c$ |
| \$10 Player | $69.5_a$ | $55.7_b$ | $50.2_c$ |

*Note:* See Table 1.

Generally, the demands reflected the reputations established (see Table 4): in the control condition, most of the offers centered around 50–50. There were 9 disagreements in each of the 50–50 and 80–20 conditions, and 15 disagreements in the control condition. Although these frequencies are not significantly different from one another ($\chi^2(2) = 2.18$), they suggest that the experimental treatment contributed to a reduction in the number of disagreements and an increase in efficiency of the bargaining, as compared to the control condition.

TABLE 3—UNAGGREGATED DATA

| | Control | | 50–50 Condition | | 20–80 Condition | |
|---|---|---|---|---|---|---|
| Trial | Outcome | Player | Outcome | Player | Outcome | Player |
| 16 | 44–52 | 1–6 | 0–0 | 1–10 | 20–80 | 1–10 |
| | 45–55 | 3–8 | 50–50 | 3–2 | 50–50 | 3–2 |
| | 0–0 | 5–10 | 0–0 | 5–4 | 0–0 | 5–4 |
| | 0–0 | 7–2 | 50–50 | 7–6 | 25–75 | 7–6 |
| | 42–58 | 9–4 | 49–51 | 9–8 | 0–0 | 9–8 |
| 17 | 47–53 | 1–10 | 50–50 | 1–8 | 22–62 | 1–2 |
| | 50–50 | 3–2 | 53–47 | 3–10 | 35–50 | 3–4 |
| | 45–50 | 5–4 | 0–0 | 5–2 | 30–70 | 5–6 |
| | 40–60 | 7–6 | 50–50 | 7–4 | 25–60 | 7–8 |
| | 40–60 | 9–8 | 45–51 | 9–6 | 20–80 | 9–10 |
| 18 | 50–50 | 1–2 | 50–50 | 1–6 | 20–79 | 1–8 |
| | 45–55 | 3–4 | 0–0 | 3–8 | 0–0 | 3–10 |
| | 0–0 | 5–6 | 50–50 | 5–10 | 30–70 | 5–2 |
| | 40–60 | 7–8 | 50–50 | 7–2 | 50–50 | 7–4 |
| | 37–63 | 9–10 | 48–52 | 9–4 | 25–70 | 9–6 |
| 19 | 45–53 | 1–4 | 0–0 | 1–4 | 20–80 | 1–6 |
| | 50–50 | 3–6 | 50–50 | 3–6 | 0–0 | 3–8 |
| | 0–0 | 5–8 | 50–50 | 5–8 | 12–88 | 5–10 |
| | 35–65 | 7–10 | 50–50 | 7–10 | 21–79 | 7–2 |
| | 46–50 | 9–2 | 49–51 | 9–2 | 25–70 | 9–4 |
| 20 | 41–59 | 1–4 | 50–50 | 1–10 | 22–65 | 1–4 |
| | 50–50 | 3–6 | 50–50 | 3–2 | 25–50 | 3–6 |
| | 35–65 | 5–8 | 50–50 | 5–4 | 25–70 | 5–8 |
| | 0–0 | 7–10 | 50–50 | 7–6 | 21–79 | 7–10 |
| | 46–50 | 9–2 | 0–0 | 9–8 | 20–80 | 9–2 |
| 21 | 34–66 | 1–10 | 50–50 | 1–2 | 20–80 | 1–6 |
| | 50–50 | 3–2 | 0–0 | 3–4 | 50–50 | 3–8 |
| | 0–0 | 5–4 | 50–50 | 5–6 | 0–0 | 5–10 |
| | 45–50 | 7–6 | 50–50 | 7–8 | 50–50 | 7–2 |
| | 35–65 | 9–8 | 49–51 | 9–10 | 32–65 | 9–4 |
| 22 | 50–50 | 1–2 | 50–50 | 1–6 | 0–0 | 1–10 |
| | 45–55 | 3–4 | 50–50 | 3–8 | 11–89 | 3–2 |
| | 0–0 | 5–6 | 50–50 | 5–10 | 33–66 | 5–4 |
| | 0–0 | 7–8 | 50–50 | 7–2 | 0–0 | 7–6 |
| | 34–66 | 9–10 | 49–51 | 9–4 | 25–65 | 9–8 |
| 23 | 40–55 | 1–6 | 50–50 | 1–4 | 21–78 | 1–2 |
| | 0–0 | 3–8 | 0–0 | 3–6 | 35–50 | 3–4 |
| | 0–0 | 5–10 | 50–50 | 5–8 | 30–70 | 5–6 |
| | 0–0 | 7–2 | 50–50 | 7–10 | 25–75 | 7–8 |
| | 0–0 | 9–4 | 50–50 | 9–2 | 0–0 | 9–10 |
| 24 | 0–0 | 1–8 | 0–0 | 1–8 | 20–80 | 1–8 |
| | 45–55 | 3–10 | 50–50 | 3–10 | 0–0 | 3–10 |
| | 50–50 | 5–2 | 50–50 | 5–2 | 24–76 | 5–2 |
| | 45–55 | 7–4 | 50–50 | 7–4 | 50–50 | 7–4 |
| | 41–58 | 9–6 | 47–53 | 9–6 | 30–70 | 9–6 |
| 25 | 50–50 | 1–2 | 50–50 | 1–2 | 20–80 | 1–10 |
| | 0–0 | 3–4 | 50–50 | 3–4 | 20–80 | 3–2 |
| | 40–60 | 5–6 | 50–50 | 5–6 | 30–70 | 5–4 |
| | 0–0 | 7–8 | 50–50 | 7–8 | 25–75 | 7–6 |
| | 34–66 | 9–10 | 49–50 | 9–10 | 25–75 | 9–8 |

*Note:* For each condition, for each trial, each line gives the outcomes for the player (with the $40 player's outcome first) and the players' number. Notice that odd numbered players are the $40 ones. For example, the first line in the control condition reads: ($40) player 1 bargained with ($10) player 6 and the outcome was 44 percent of the lottery tickets for 1 and 52 percent of the lottery tickets for player 6.

TABLE 4—REPUTATION[a]

| | | Control Condition | | | 50–50 Condition | | | 20–80 Condition | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Player | Trial | First Demand | Second Demand | Outcome | First Demand | Second Demand | Outcome | First Demand | Second Demand | Outcome |
| 1 | 11 | 62 | 35 | 35 | 50 | 50 | 50 | 20 | 20 | 20 |
| | 12 | 55 | 55 | 55 | 51 | 51 | 0 | 20 | 20 | 20 |
| | 13 | 55 | 51 | 51 | 50 | 50 | 50 | 23 | 20 | 20 |
| | 14 | 47 | 47 | 0 | 51 | 51 | 0 | 23 | 20 | 20 |
| | 15 | 55 | 35 | 35 | 50 | 50 | 50 | 23 | 20 | 20 |
| 2 | 11 | 49 | 49 | 49 | 35 | 51 | 51 | 84 | 79 | 79 |
| | 12 | 49 | 49 | 49 | 55 | 52 | 52 | 77 | 77 | 77 |
| | 13 | 49 | 49 | 49 | 55 | 52 | 52 | 86 | 78 | 78 |
| | 14 | 49 | 49 | 49 | 55 | 50 | 50 | 92 | 79 | 79 |
| | 15 | 55 | 50 | 50 | 55 | 53 | 53 | 80 | 80 | 80 |
| 3 | 11 | 50 | 50 | 50 | 50 | 50 | 50 | 70 | 21 | 21 |
| | 12 | 50 | 50 | 0 | 55 | 55 | 0 | 80 | 24 | 24 |
| | 13 | 50 | 50 | 50 | 50 | 50 | 50 | 100 | 20 | 20 |
| | 14 | 50 | 40 | 40 | 51 | 51 | 0 | 50 | 25 | 25 |
| | 15 | 50 | 50 | 50 | 50 | 50 | 50 | 65 | 21 | 21 |
| 4 | 11 | 68 | 68 | 68 | 55 | 51 | 51 | 80 | 79 | 79 |
| | 12 | 62 | 62 | 62 | 55 | 51 | 51 | 65 | 65 | 65 |
| | 13 | 85 | 50 | 50 | 80 | 52 | 52 | 100 | 75 | 75 |
| | 14 | 65 | 65 | 0 | 80 | 80 | 0 | 100 | 78 | 78 |
| | 15 | 56 | 56 | 50 | 80 | 52 | 52 | 100 | 79 | 79 |
| 5 | 11 | 51 | 32 | 32 | 55 | 55 | 0 | 29 | 29 | 0 |
| | 12 | 51 | 51 | 0 | 50 | 50 | 50 | 70 | 70 | 0 |
| | 13 | 51 | 51 | 0 | 72 | 27 | 27 | 40 | 40 | 0 |
| | 14 | 50 | 40 | 40 | 60 | 29 | 29 | 35 | 35 | 0 |
| | 15 | 50 | 45 | 45 | 50 | 50 | 50 | 46 | 20 | 20 |
| 6 | 11 | 75 | 75 | 0 | 54 | 53 | 53 | 82 | 82 | 0 |
| | 12 | 75 | 45 | 45 | 53 | 53 | 0 | 80 | 80 | 80 |
| | 13 | 75 | 75 | 0 | 52 | 52 | 52 | 80 | 80 | 80 |
| | 14 | 60 | 50 | 50 | 65 | 52 | 52 | 80 | 80 | 80 |
| | 15 | 60 | 56 | 56 | 59 | 52 | 52 | 80 | 80 | 80 |
| 7 | 11 | 90 | 90 | 0 | 50 | 50 | 50 | 27 | 20 | 20 |
| | 12 | 50 | 50 | 50 | 50 | 50 | 50 | 25 | 23 | 23 |
| | 13 | 55 | 40 | 40 | 51 | 51 | 0 | 27 | 20 | 20 |
| | 14 | 75 | 75 | 0 | 50 | 50 | 50 | 26 | 22 | 22 |
| | 15 | 55 | 44 | 44 | 50 | 50 | 50 | 30 | 20 | 20 |
| 8 | 11 | 65 | 65 | 0 | 51 | 51 | 51 | 85 | 78 | 78 |
| | 12 | 60 | 60 | 0 | 51 | 51 | 51 | 80 | 74 | 74 |
| | 13 | 60 | 60 | 60 | 49 | 49 | 49 | 90 | 78 | 78 |
| | 14 | 60 | 60 | 60 | 50 | 50 | 50 | 80 | 80 | 80 |
| | 15 | 65 | 65 | 65 | 51 | 51 | 0 | 80 | 79 | 79 |
| 9 | 11 | 45 | 45 | 0 | 52 | 52 | 0 | 30 | 20 | 20 |
| | 12 | 44 | 38 | 38 | 47 | 28 | 28 | 35 | 20 | 20 |
| | 13 | 43 | 43 | 0 | 49 | 49 | 49 | 35 | 24 | 24 |
| | 14 | 50 | 50 | 50 | 49 | 49 | 49 | 55 | 20 | 20 |
| | 15 | 44 | 40 | 40 | 49 | 49 | 49 | 33 | 20 | 20 |
| 10 | 11 | 65 | 65 | 65 | 54 | 53 | 53 | 82 | 82 | 0 |
| | 12 | 70 | 70 | 0 | 54 | 52 | 52 | 80 | 80 | 80 |
| | 13 | 76 | 76 | 0 | 52 | 52 | 52 | 80 | 80 | 80 |
| | 14 | 66 | 66 | 0 | 53 | 53 | 53 | 81 | 81 | 0 |
| | 15 | 69 | 50 | 50 | 54 | 51 | 51 | 80 | 80 | 80 |

[a] Odd numbered players had a $40 prize, even numbered players had a $10 prize.

## V. Discussion

The results of this experiment provide strong support for the hypothesis that the outcome of the bargaining is influenced not only by the preferences and strategic options of the bargainers, but also by their expectations. By manipulating these expectations, it proved possible to consistently produce outcomes which differed significantly from those observed when the players' expectations were not manipulated. The fact that there was no significant difference over trials indicates that these expectations were self-reinforcing, so that the outcomes which resulted in the experimental conditions were stable and self-sustaining. Players who expected their opponents to expect equal payoffs (or an equal division of lottery tickets) continued to meet such opponents, and consequently those expectations were reinforced.[5]

These results also lend support to the hypothesis put forward in Section I to explain the results of the previous experiments. Since those experiments involved no artificial manipulation of the bargainers' expectations, it must be that the (common) expectations which bargainers formed when different kinds of information were available resulted in large measure from their previous experience. This suggests that there must be many kinds of potential conflict in which individuals have common expectations which permit them to efficiently reach agreement. The benefits to a society of fostering such common expectations are obvious, since otherwise many bargaining situations would end in disagreement.

The data from the present experiment suggest that individuals entered the experiment with more or less mutually consistent prior expectations about what kinds of agreements would result, and that they updated these expectations in response to their experience in the experiment. These prior expectations, as reflected in the control condition, yielded outcomes closer to 50–50 than to 20–80. It is perhaps for this reason that the outcomes in the 50–50 condition showed so much smaller variance than those in the 20–80 condition.

These results have several implications for the development of the theory of bargaining. The most striking is that it may be necessary to incorporate the expectations of the bargainers into any description (or definition) of equilibrium outcomes, and that there may in general be multiple equilibria supported by different sets of mutually consistent expectations. Some models consistent with this suggestion have already been explored, and the notion that agents' beliefs play a role in determining outcomes is not a new one.[6] However, the results of this experiment suggest that, because of the role which agents' beliefs play in determining the outcome of bargaining, it may be necessary to look to models in which at least some elements of the expectations which players bring to the bargaining will be empirically determined exogenous variables.

---

[5] The experimental conditions can be interpreted as having changed the *strategic risk posture* of the players, as measured by the certain payoff which they would regard as equivalent to the (risky) opportunity to engage in the bargaining. The notion of strategic risk posture (introduced in Roth, 1977a,b), and studied in the context of bargaining games in Roth (1978, 1979) plays a role parallel to ordinary risk posture in determining an individual's utility for engaging in a game. The role of ordinary risk posture in bargaining models is explored in Roth and Uriel Rothblum (1982).

[6] An interesting paper in which a general definition of equilibrium in games is proposed which explicitly involves certain beliefs of the players is Kreps and Wilson (1982a).

## REFERENCES

**Aumann, Robert J.**, "Agreeing to Disagree," *The Annals of Statistics*, November 1976, *4*, 1236–39.

**Edgeworth, F. Y.**, *Mathematical Psychics*, London: Routledge and Kegan Paul, 1881.

**Harsanyi, John C.**, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge: Cambridge University Press, 1977.

**Kreps, David M. et al.**, "Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma," *Journal of Economic Theory*, August 1982, *27*, 245–52.

______ **and Wilson, Robert**, (1982a) "Sequential Equilibria," *Econometrica*, September

1982, *50*, 863–94.

______ and ______, (1982b) "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, *27*, 253–79.

**Lewis, David K.**, *Convention: A Philosophical Study*, Cambridge: Harvard University Press, 1969.

**Milgrom, Paul**, "An Axiomatic Characterization of Common Knowledge," *Econometrica*, January 1981, *49*, 219–21.

______ **and Roberts, John**, "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, August 1982, *27*, 280–312.

**Nash, John F.**, "The Bargaining Problem," *Econometrica*, April 1950, *18*, 155–62.

______, "Non Cooperative Games," *Annals of Mathematics*, September 1951, *54*, 286–95.

**Rosenthal, Robert W.**, "Sequences of Games With Varying Opponents," *Econometrica*, November 1979, *47*, 1353–66.

______ **and Landau, Henry J.**, "A Game-Theoretic Analysis of Bargaining With Reputations," *Journal of Mathematical Psychology*, December 1979, *20*, 233–55.

**Roth, Alvin E.**, (1977a) "The Shapley Value as a von Neumann-Morgenstern Utility Function," *Econometrica*, April 1977, *45*, 657–64.

______, (1977b) "Bargaining Ability, The Utility of Playing a Game, and Models of Coalition Formation," *Journal of Mathematical Psychology*, October 1977, *16*, 153–60.

______, "The Nash Solution and the Utility of Bargaining," *Econometrica*, May 1978, *46*, 587–94.

______, *Axiomatic Models of Bargaining*, Berlin: Springer, 1979.

______ **and Malouf, Michael W. K.**, "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, November 1979, *86*, 574–94.

______, ______, **and Murnighan, J. Keith**, "Sociological versus Strategic Factors in Bargaining," *Journal of Economic Behavior and Organization*, June 1981, *2*, 153–77.

______ **and Murnighan, J. Keith**, "The Role of Information in Bargaining: An Experimental Study," *Econometrica*, September 1982, *50*, 1123–42.

______ **and Rothblum, Uriel G.**, "Risk Aversion and Nash's Solution for Bargaining Games with Risky Outcomes," *Econometrica*, May 1982, *50*, 639–47.

# Asymmetric Price Rigidity and Inflationary Bias

*By* TIMUR KURAN*

When prices adjust in response to market disturbances, they usually do so in discrete jumps. Even when market conditions are changing continuously, prices tend to remain fixed for substantial lengths of time. The commodities fitting this general pattern include most retail goods, services, and intermediate materials.[1]

While it is widely recognized that this pricing pattern holds for both upward and downward adjustments, there is no consensus as to whether these are mirror images for given individual firms. Some scholars believe that upward adjustments are generally more pronounced in magnitude or frequency than downward adjustments. Robert Solow (1980), for instance, believes on the basis of casual observations of some macroeconomic evidence, that commodity prices are sticky downward.

Many others, however, view the issue as a completely open question. Phillip Cagan (1979) has documented that prices in the United States tend to rise more in response to increases in demand than they fall in response to corresponding decreases in demand; but he is careful to point out that our present knowledge does not permit us to attribute his evidence to biases in the adjustment behavior of individual firms. Arthur Okun's (1981) research, in contrast, does not point toward an upward bias in prices, yet he, too, agrees that the issue is far from closed.[2]

It is quite possible, of course, for the adjustment policies of different types of firms to be asymmetric in opposite directions, and for some firms to display perfectly symmetric behavior. It is also possible for a given firm to exhibit asymmetric pricing behavior in response to some kinds of disturbances but not others. If indeed there are firms in the economy which display asymmetric behavior in one situation or another, one ought to be able to establish this both empirically and analytically.

This paper examines the adjustment policies of one particular class of business firms: profit-maximizing monopolistic firms with nondecreasing marginal cost functions and nonincreasing elasticity demand functions. The specific disturbances considered in the analysis are changes in the general price level.

I show that firms in this class which incur fixed-price-adjustment costs are likely to change their prices less often under deflation than they would under an equivalent inflation. I also demonstrate that the price-adjustment policies of these firms can be biased upward, in the sense that under given inflationary expectations the firms raise their prices more than they would lower them were they to have symmetric deflationary expectations. Finally, I establish that inflationary uncertainty about *any* core rate of inflation can heighten the upward bias in the firms' adjustment policies.

The intuitive explanation for all these findings is that for a given firm the opportunity cost in any period of charging a suboptimal real price is not symmetric around the firm's optimal price. Money illusion is not the cause of this basic asymmetry: it exists even if the firm, its consumers, and sellers of its inputs all base their decisions on real magnitudes.

---

*Assistant professor of economics, University of Southern California. I am indebted for helpful criticisms and suggestions to Kenneth Arrow, Paul Evans, David Starrett, and an anonymous referee. Responsibility for any errors or omissions is, of course, entirely mine.

[1] The main exception is organized exchanges for raw materials.

[2] Cagan's hesitation stems partly from arguments that secret discounting imparts flexibility to actual transaction prices when the quoted prices are rigid. He does concede, however, that even the important empirical study of George Stigler and James Kindahl (1970), undertaken to demonstrate that prices are substantially more flexible than they appear in official statistics, fails to settle the issue. As for Okun's doubts, they appear to be linked to the administered price thesis expounded by Gardiner Means and others (1975). Robert Gordon (1981) provides a splendid summary of the macroeconomic evidence and a survey of related theoretical arguments.

In the model presented, the firm's cumulative price adjustment cost is assumed to be invariant to the direction and size of adjustment. This assumption is realistic when the cost arises entirely from the price decision process and the dissemination of price information.[3] For any actual firm, though, there exist various other costs which may differ between upward and downward adjustments,[4] or with the size of adjustment, and asymmetries could arise in the firm's price-adjustment policy for this reason alone. The rationale for assuming that the cumulative cost is fixed is not to deny that it may differ between upward and downward movements for an individual firm, but rather to demonstrate that asymmetries can exist in the firm's price-adjustment policy even when the adjustment cost is fixed.

An important simplifying assumption in the model is that the firm has a two-period planning horizon. In a more complete formulation, the horizon would depend explicitly on the costs of developing and disseminating price policies as well as on the degree of uncertainty concerning demand and the general price level.[5] In practice, since the future is uncertain and planning is costly, firms use finite planning horizons, the lengths of which vary widely. Each firm determines the length of its horizon according to such characteristics as market structure, the quality of information about market conditions, planning costs, and industry rules of thumb. The two-period formulation presented here should be regarded as a simplification of the finite-horizon planning model which captures the essential aspects of the problem.

A related simplifying assumption is that the firm's horizon is fixed, in that successive horizons do not overlap. This means that the firm cannot modify a price decision made at the beginning of a horizon. In practice, even firms which issue catalogs at pre-announced intervals reserve the right to change their prices without notice. Due to the heavy costs involved, however, they prefer to exercise this right only in the event of a great surprise. Mail-order firms, for instance, issue price catalogs at intervals which depend on their expectations about inflation.[6] While their predictions of the price level frequently turn out to be wrong, they do not revise their prices ahead of schedule, unless the prediction error is drastically large.

Section I contains a formulation of the basic model in which the firm has a constant elasticity demand function and a constant marginal cost function, and is assumed to know with certainty the proportional change in the price level between the two periods in its planning horizon. The section proceeds to determine the range of the inflation factor for which price adjustment is unprofitable. Section II then shows that this range is generally not symmetric around one. Section III establishes that the firm's optimal price-adjustment policy is biased upward, and that uncertainty would compound the inflationary bias. The formal results are generalized considerably in Section IV to include any nonincreasing elasticity demand function and any nondecreasing marginal cost function. Section V summarizes the results and touches upon their potential macroeconomic implications.

---

[3] Decision and dissemination costs are incurred through activities such as detecting the need for a change, altering bookkeeping entries, printing and mailing new catalogs, changing price labels, and readjusting coin vending machines.

[4] For many firms, as Okun has discussed in detail, customers' need for dependability imposes costs for upward but not downward price changes. The same asymmetry exists for regulated firms which can gain approval for price increases only through costly lobbying efforts, but can lower their prices at will. An opposite asymmetry is present in oligopolistic industries where a price decrease, but not an increase, carries with it the risk of a costly price war.

[5] See Michael Intriligator and Eytan Sheshinski (1979) for a general analysis of the horizon decision in economic planning. They show that the optimal planning horizon is infinite only in the hypothetical case where the firm knows with perfect certainty demand conditions and the price level for all time periods.

[6] For some casual evidence from the United States, see the *Wall Street Journal* news briefs assembled by Sheshinski and Yoram Weiss (1977).

## I. The Model

### A. *Profit Functions*

Consider a profit-maximizing monopolist with a two-period planning horizon. This firm's demand in period 1 is given by

$$(1)\qquad x_1 = ap_1^{-b}, \qquad a>0,\ \ b>1,$$

where $a$ specifies the market size in period 1, $b$ is the price elasticity of demand, and $p$ denotes nominal price per unit of output. The general price level is normalized to be one in the first period, making the nominal and real prices equal. In period 2, the firm expects its demand to be

$$(2)\qquad x_2 = a\theta(p_2/\mu)^{-b}, \qquad 0 \leqslant \theta < \infty, \quad 0<\mu<\infty,$$

where $\theta$ and $\mu$ are proportional indexes of the changes in its market size and the price level, respectively.[7] When $\theta$ and $\mu$ are both equal to one, the firm expects neither its market size nor the price level to change relative to their values in period 1. A value of $\theta$ less than one (greater than one) implies that the firm expects its market to shrink (expand) in period 2. When $\mu$ is less than (greater than) one, the firm expects the price level to fall (rise). For the time being, it will be assumed that the firm knows $\theta$ and $\mu$ with certainty. It is important to emphasize that money illusion is absent from this demand system since the homogeneity postulate of neoclassical economics is satisfied.

For each unit of output, the firm incurs a constant positive real cost denoted by $c$. In nominal terms, therefore, cost per unit is $c$ in period 1 and $c\mu$ in period 2. If the firm chooses to charge different nominal prices in each period, it also incurs a real adjustment cost, $\eta$, in period 2.[8] Assuming there is no capacity constraint, the firm's nominal profits in each period are given by

$$(3)\qquad \Pi_1(p_1) = x_1(p_1 - c);$$

$$(4)\qquad \Pi_2(p_2) = x_2(p_2 - c\mu) - \delta(\eta\mu),$$

where $\delta$ is a Kronecker *delta* which takes the value 0 if $p_1 = p_2$, and the value 1 otherwise. Note that nominal changes are assumed to be costly even when no real change is involved. This is a natural assumption since almost all firms set their prices in nominal terms. Even in high-inflation countries where wage and bond indexation are common, very few prices, wholesale or retail, are linked to the general price level.[9]

The assumption that real input costs are impervious to changes in the price level has a significant bearing on the results to be presented, so it must be justified before proceeding. There are two rationales for this assumption, one involving material costs and the other labor costs. The material component of costs could conceivably fluctuate over time because the suppliers themselves incur price adjustment costs. Considering, however, that the production of almost any commodity requires multiple material inputs and that price adjustment patterns are imperfectly synchronized across suppliers, it appears that the cumulative real cost of material inputs would not in general be seriously affected by changes in the price level. The labor component of costs, on the other hand, is not affected by inflation because wages are often implicitly, and sometimes explicitly, linked to the price level.[10]

### B. *Alternative Price Policies*

At the beginning of period 1, the firm must decide whether to charge the same or different nominal prices in the two periods.

---

[7] Implicit in this formulation is the assumption that the firm is small enough to ignore its own influence on the price level.

[8] The cost of choosing a price in period 1 is treated as a fixed cost and hence disregarded.

[9] In an economy where firms' prices were indexed, price level movements would generate automatic and costless changes in nominal prices. In such an economy, firms would incur costs only for real price changes.

[10] See Okun (chs. 2, 3, and 6) for an analysis of why wage indexation is so prevalent.

To determine the profit-maximizing sequence of prices, the firm calculates and then compares the present discounted value of total real profits under each of these two price policies. The real discount factor is $\rho = 1/(1+r)$, where $r$ denotes the real interest rate.

If the firm chooses to pursue a one-price policy, the optimal nominal price is that which maximizes its total discounted real profits

$$\Pi(p) = \Pi_1(p) + \rho\mu^{-1}\Pi_2(p), \tag{5}$$

where $p = p_1 = p_2$, and $\Pi_2(p)$ is given by (4) with $\delta = 0$. Taking the first-order condition of (5) and rearranging terms gives the optimal $p$ as

$$p^* = P\frac{(1+\rho\theta\mu^b)}{(1+\rho\theta\mu^{b-1})}, \tag{6}$$

where $P = bc/(b-1)$. Substituting (6) into (5) yields the firm's maximum real profit as

$$\Pi^* = \phi\frac{(1+\rho\theta\mu^{b-1})^b}{(1+\rho\theta\mu^b)^{b-1}}, \tag{7}$$

where $\phi = aP^{-b}(P-c)$.

The firm's other option is to pursue a two-price policy. In this case, its maximand takes the form

$$\Pi(p_1, p_2) = \Pi_1(p_1) + \rho\mu^{-1}\Pi_2(p_2), \tag{8}$$

where $\Pi_2(p_2)$ is given by (4) with $\delta = 1$. The first-order conditions of (8) give

$$p_1^{**} = P; \tag{9}$$

$$p_2^{**} = P\mu, \tag{10}$$

which are the monopoly prices in each period. Note that $p_1$ and $p_2$ are independent of $a$ and $\theta$. Whereas the optimal one-price policy almost always depends on the relative market sizes in the two periods,[11] here $p_1$ and $p_2$ are invariant with respect to the market size parameters. By substituting the optimal prices into (7), we obtain the maximum level of discounted real profits as

$$\Pi^{**} = \phi(1+\rho\theta) - \rho\eta. \tag{11}$$

### C. *Optimal Price Policy*

Now let us define the function

$$G(\mu, \theta) \equiv \Pi^* - \Pi^{**}. \tag{12}$$

Given a pair of values for $\mu$ and $\theta$, the sign of $G$ determines the firm's optimal price regime. The one-price policy is optimal when $G > 0$, and the two-price policy is optimal when $G < 0$. In cases where $G = 0$, the firm is indifferent between the two regimes, but to facilitate exposition of the results, it is assumed that the firm favors the one-price policy.

To establish the properties of the function $G$, which is depicted in Figure 1, it is useful to begin by examining its slope as $\mu$ varies between 0 and $\infty$. Differentiating $G$ with respect to $\mu$ and rearranging terms reveals that the function $G$ monotonically increases as $\mu$ varies between 0 and 1, reaches a maximum at $\mu = 1$, and monotonically decreases as $\mu$ increases beyond 1. By substituting $\mu = 1$ into (12), one can see that the maximum value of $G$ is $\rho\eta$. Intuition bears this out: when $\mu = 1$ the optimal price and output levels are the same under the two alternative price policies, but real profits are lower under the two-price policy because of the adjustment cost. As $\mu$ moves away from 1 in either direction, $G$ takes on values less than $\rho\eta$ because $\Pi^*$ monotonically declines.

It follows that the equation $G(\mu, \theta) = 0$ has a maximum of two roots in $\mu$. These critical values, when they exist, will be denoted by $\underline{\mu}(\theta)$ and $\bar{\mu}(\theta)$, designating, respectively, the lowest and highest values of $\mu$ for which the one-price policy is optimal. Only if $\mu$ falls outside this closed interval is it optimal to charge different nominal prices in each period.

---

[11] The optimal one-price policy is independent of $\theta$ only when $\mu = 1$. In this special case, the vector of optimal prices is the same under each policy, but profits are lower by the amount $\rho\eta$ under the one- than the two-price policy.

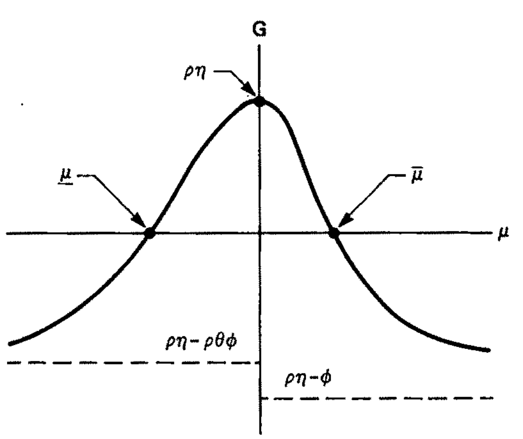


FIGURE 1. THE FUNCTION $G$ AND ITS ASYMPTOTES FOR GIVEN $\theta$ (SEMILOGARITHMIC SCALE)

## II. Analysis of the Critical Values

### A. *Existence*

The conditions under which $\underline{\mu}$ and $\bar{\mu}$ exist can be determined by examining the behavior of $G$ as $\mu$ approaches 0 and $\infty$. The limit of $G$ as $\mu$ approaches 0 is $\rho\eta-\rho\phi\theta$, implying that $\underline{\mu}$ exists if and only if $\eta \leqslant \phi\theta$. The limit as $\mu$ approaches $\infty$ is $\rho\eta-\phi$, implying that $\bar{\mu}$ exists if and only if $\eta \leqslant \phi/\rho$. Thus, a sufficient condition for the existence of both critical values is that $\eta$ fall short of the minimum of $\phi\theta$ and $\phi/\rho$. Since the firm is likely to have chosen its pricing period sufficiently long to make this condition hold, it shall be assumed in the following analysis that $\underline{\mu}$ and $\bar{\mu}$ do exist.

### B. *Asymmetry*

Let us begin by examining the properties of the set which consists of the ordered pairs $(\mu,\theta)$ satisfying the equation $G(\mu,\theta)=0$. This equation implicitly defines $\underline{\mu}$ and $\bar{\mu}$ as a function of $\theta$. An important initial question is whether $\underline{\mu}(\theta)$ and $\bar{\mu}(\theta)$ are generally symmetric around 1. The relevant concept of symmetry is, of course, the geometric one, since $\mu$ represents *proportional* changes in the price level. By definition, $\underline{\mu}$ and $\bar{\mu}$ are symmetric if and only if their geometric mean $\hat{\mu}$ is equal to 1. The following proposition establishes the magnitude of $\hat{\mu}$ in relation to 1:[12]

PROPOSITION 1: $\hat{\mu}(\theta) \gtreqless$ *if and only if* $\theta \gtreqless 1/\rho$.

It follows from this proposition, since $1/\rho = 1+r$, that $\underline{\mu}$ and $\bar{\mu}$ are symmetric around 1 when $\theta-1=r$, where $\theta-1$ can be interpreted as the growth rate of the size of the market.[13] When the market growth rate exceeds the interest rate, the one-price policy remains optimal for proportionally larger increases in the price level than decreases. In contrast, when the growth rate is lower than the interest rate, the one-price policy stays optimal for larger decreases than increases.

The functions $\underline{\mu}(\theta)$, $\bar{\mu}(\theta)$, and $\hat{\mu}(\theta)$ are illustrated in Figure 2. In this figure the region bounded by and including the functions $\underline{\mu}(\theta)$ and $\bar{\mu}(\theta)$ is the locus of ordered pairs $(\mu,\theta)$ for which the one-price policy is optimal. The limits of $\underline{\mu}(\theta)$ and $\bar{\mu}(\theta)$ as $\theta$ tends to infinity are denoted by $\underline{\mu}\infty$ and $\bar{\mu}\infty$, and their geometric mean, which by Proposition 1 is necessarily greater than one, is denoted by $\hat{\mu}\infty$.[14]

The intuitive explanation for the proposition centers around two properties of the model: 1) the growth rate and the interest rate affect the valuation of first- vs. second-period profits in opposite ways. This valuation is an increasing function of the interest rate, but a decreasing function of the growth rate. The two effects balance each other out only when $\theta=1/\rho$. 2) The firm's profits in any given period are greater for an upward displacement from the monopoly

---

[12] The proof contains two major steps. First, it must be shown that $G(\mu,\theta) \gtreqless G(1/\mu,\theta)$ as $\theta \gtreqless 1/\rho$, which implies, since $G(\bar{\mu},\theta)=0$ by the definition of $\bar{\mu}$, that $G(1/\bar{\mu},\theta) \lesseqgtr 0$ as $\theta \gtreqless 1/\rho$. Application of some of the properties of $G$ outlined in Section I is then sufficient to establish the proposition.

[13] In aggregative growth theory, this condition describes the Golden Rule path along which consumption per capita is maximized; but the Golden Rule and the symmetry condition in the present model arise from two entirely different sets of economic assumptions.

[14] Figures 1 and 2 are complementary. While the former indicates the value of $G$ as a function of $\mu$ for fixed $\theta$, the latter shows $\mu$ as a function of $\theta$ for one particular level curve of $G$, namely that where $G=0$.

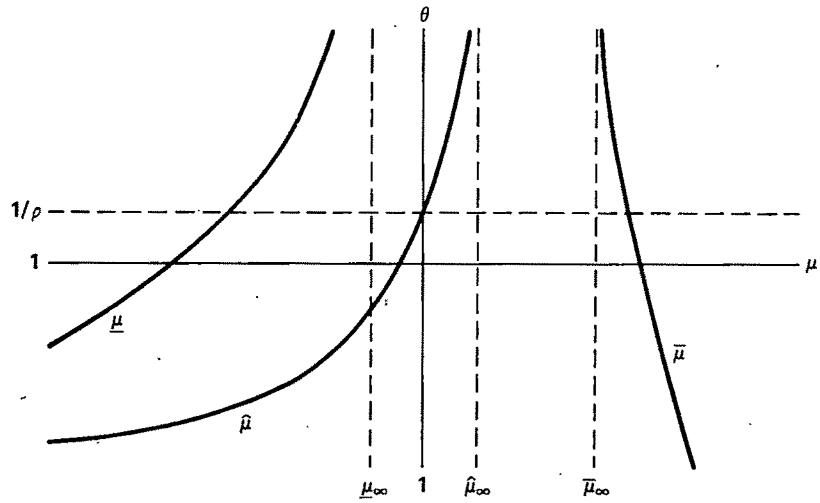


FIGURE 2. CRITICAL VALUES OF $\mu$ AS A FUNCTION OF $\theta$ (LOGARITHMIC SCALE)

price than for a proportionally equivalent downward displacement.[15] The source of this asymmetry can be traced (as discussed more fully in Section IV) to the facts that the price elasticity of demand is nonincreasing, and that real marginal cost is nondecreasing.

Of the two asymmetric cases, the one where the growth rate is less than the interest rate is more likely to be observed in any actual economy, so let us examine it in some detail. In this case, the valuation of first-period profits exceeds that of second-period profits. Under inflation, therefore, when the firm is pursuing a one-price policy, its real price is too high in the period weighted more heavily and too low in the other period. Under deflation, in contrast, its real price is too low in the period weighted more heavily and too high in the other period. Thus, if the profit function were symmetric, the firm would want, under inflation, to charge a price closer to the lower of the two monopoly prices. Under deflation, on the other hand, it would want to charge a price closer to the higher of the two monopoly prices.

But the profit function is *not* symmetric, and this requires the firm to charge a price closer to the higher of the two monopoly prices. Under inflation the asymmetry pulls the optimal price toward the period weighted less heavily, while under deflation the pull is toward the period weighted more heavily. So the reason why the firm switches to a two-price policy more readily under inflation is that the two forces, one due to discounted growth and the other to the asymmetry in the profit function, are counteracting.

Naturally, if the growth rate were greater than the interest rate the forces denoted above would be counteracting under deflation and reinforcing under inflation. The firm would then switch to a two-price policy more readily under deflation.

These results show that in the presence of an adjustment cost price rigidity can be greater in one direction than the other. In the case where the interest rate exceeds the growth rate, for instance, prices would change more often under inflation than deflation.

## III. Inflationary Bias

### A. *Comparison of Responses to Downward and Upward Price Level Changes*

While the discussion thus far has analyzed the conditions under which each of the firm's alternative pricing policies is optimal, it is necessary to examine now whether the firm responds differently to equivalent doses of inflation and deflation in its price-setting strategy. The analysis involves comparing $\tilde{p}$, the geometric average of the first- and second-period nominal prices, for proportionally equal price level changes in either direction.[16] Note that if $\mu$ were sufficiently high or low to make the firm pursue a two-price policy, $\tilde{p}$ would equal $P\mu^{1/2}$.

---

[15] The asymmetry is most easily seen by comparing the two extreme displacements, $p=\infty$ and $p=0$. Raising the firm's price to $\infty$ reduces profits to 0, but lowering it to 0 reduces profits to $-\infty$.

If the firm's price is more responsive to a rise in the price level than to a proportionally equal fall, we shall say that its price adjustment policy exhibits inflationary bias. Thus, for a given $\mu > 1$ the firm's price-adjustment policy is biased toward inflation if $\hat{p}(\mu)$, the geometric average of $\tilde{p}(\mu)$ and $\tilde{p}(1/\mu)$, exceeds $P$. The firm's adjustment policy is unbiased if $\tilde{p}(\mu)$ and $\tilde{p}(1/\mu)$ differ from $P$ in the same proportion, that is, if $\hat{p}(\mu)$ equals $P$. The following proposition establishes how $\hat{p}$ varies according to the level of $\mu$.[17]

PROPOSITION 2: *If* $1 < \mu \leqslant max[\bar{\mu}, 1/\underline{\mu}]$, *then* $\hat{p}(\mu) > P$. *Otherwise*, $\hat{p}(\mu) = P$.

This proposition indicates that the firm's responses to inflation and deflation are symmetric if $\mu > \bar{\mu}$ and $1/\mu < \underline{\mu}$. The firm is more responsive to a rise in the price level than to a fall if either $\mu$ or $1/\mu$ (or both) are located in the one-price adjustment zone. Note that the proposition holds irrespective of the size of $\theta$ in relation to $1/\rho$. In particular, when the growth rate and the interest rate are equal, so that $\mu$ and $\bar{\mu}$ are symmetric, $\hat{p}(\mu)$ still exceeds $P$, provided $\mu$ or $1/\mu$ is in the one-price adjustment zone.

The source of inflationary bias is the already discussed asymmetry in the profit function. The firm pursues an asymmetric pricing policy to compensate for the asymmetry in profit losses from charging too high or too low a real price.

### B. *Response to Price Level Uncertainty*

Another possible manifestation of inflationary bias is the firm's response to uncertainty. Since the firm's profit function under a one-price policy is nonlinear in the logarithm of its real price, two equiprobable values of $\mu$ that are symmetric in logarithms around a given core value of $\mu$ need not exert mutually offsetting gravitational pulls on $p^*$, the optimal price. So I shall now explore whether an appropriately defined change in the degree of uncertainty about an arbitrary core rate of inflation can impart an upward bias to an individual firm's price adjustment policy.

To begin, let us drop the assumption that the firm knows $\mu$ with certainty, and postulate instead that only the distribution of $\mu$ is known. With this generalization, the first-order condition of (5) takes the form

$$p^*P^{-1} = \frac{1+\rho\theta E(\mu^b)}{1+\rho\theta E(\mu^{b-1})}, \tag{13}$$

where $E$ is the expectation operator. How a change in any given distribution affects $p^*$ depends, of course, on the nature of the change. Since $\mu$ is a proportional parameter, and more importantly, since arithmetic distributions of prices are in general heavily skewed to the right, it is meaningful to consider geometric-mean-preserving spreads (*GMPS*).[18]

The effect of a *GMPS* on $p^*$ can be measured by its effect on the ratio on the right-hand side of (13), denoted by $R$. If $R$ were to rise in response to a *GMPS*, then, if the first-order condition is to be maintained, $p^*$ would have to rise, too. How $R$ is affected by a *GMPS* can be deduced from the curvature of $R$, considered as a function of $log\mu$, in the range of the spread. A sufficient condition for $R$ to rise in response to a *GMPS* is that $R$ be strictly convex with respect to $log\mu$ throughout the range of the spread. Unfortunately, the second derivative of $R$ with respect to $log\mu$ is too complex to be evaluated analytically. Numerical analysis strongly suggests, however, that any *GMPS* would

---

[16] The appropriate comparison involves geometric averages because $\mu$ is a proportional parameter.

[17] The proof involves straightforward computation of the four possible cases: (i) $\mu \leqslant \bar{\mu}$ and $1/\mu \geqslant \underline{\mu}$, (ii) $\mu \leqslant \bar{\mu}$ and $1/\mu < \underline{\mu}$, (iii) $\mu > \bar{\mu}$ and $1/\mu \geqslant \underline{\mu}$, and (iv) $\mu > \bar{\mu}$ and $1/\mu < \underline{\mu}$.

[18] A *GMPS* of a given distribution is obtained by performing an arithmetic-mean-preserving spread on the logarithm of the distribution. See Michael Rothschild and Joseph Stiglitz (1970) for a detailed theoretical analysis of mean-preserving spreads.

TABLE 1—COMPARISON OF OPTIMAL PRICES UNDER UNCERTAINTY

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| *Log* Mean | 1.25 | 1.25 | 2.00 | 2.00 | 1.25 | 1.25 | 2.00 | 2.00 |
| *Log* Variance | 0.006 | 0.006 | 0.006 | 0.006 | 0.033 | 0.033 | 0.033 | 0.033 |
| $\rho\theta$ | 0.8 | 1.0 | 0.8 | 1.0 | 0.8 | 1.0 | 0.8 | 1.0 |
| $b=2$ | 0.5 | 0.6 | 0.7 | 0.7 | 2.9 | 3.1 | 4.0 | 4.2 |
| $b=5$ | 2.1 | 2.2 | 2.7 | 2.7 | 12.3 | 12.8 | 15.0 | 14.9 |
| $b=10$ | 5.8 | 5.8 | 5.7 | 5.7 | 27.5 | 27.3 | 26.3 | 26.3 |
| $b=20$ | 11.8 | 11.7 | 11.6 | 11.6 | 35.7 | 35.6 | 35.2 | 35.2 |

*Note:* The entries stand for percentage differences: columns (1)–(4) are based on the same *GMPS*; a second spread was performed on the resulting distribution to obtain the distribution underlying columns (5)–(8).

raise $R$ and hence $p^*$; a few examples will be presented later.

There is one type of geometric-mean-preserving spread for which the argument can be substantiated analytically. This is a uniform expansion of the *log* of the distribution around its geometric mean $\mu^*$. An expansion of degree $\alpha$ can be written as

$$(14)\qquad log\,\mu(\alpha)=log\,\mu^*+(1+\alpha)(log\,\mu-log\,\mu^*).$$

The following proposition is concerned with the effect of such an expansion.

PROPOSITION 3: *A geometric-mean-preserving uniform expansion of the distribution of* $\mu$ *raises the firm's optimal price under a one-price policy. The price rise increases with the degree of expansion.*[19]

This proposition demonstrates that uncertainty about the price level is a source of inflationary bias in that it leads the firm to charge higher prices. The explanation is again that the profit function is asymmetric in the real price: effective cost indexation, coupled with the fact that the price elasticity of demand is nonincreasing, causes the firm to be penalized more severely for underestimating the second-period price level than for overestimating it.

Some numerical examples of the effect of uncertainty on the firm's optimal price are presented in Table 1. The entries in this table show, for various parameter values and distributional assumptions, the extent to which uncertainty causes the optimal price under a one-price policy to rise above the certainty price. There are four values of $b$ shown which correspond to widely different levels of market power. At one extreme, when $b=2$, the firm's monopoly price is 100 percent higher than marginal cost; at the other, it is only about 5 percent higher.[20]

The results in Table 1 suggest that the more elastic the firm's demand, the greater the effect of uncertainty on its optimal price.[21] The explanation for this relationship is that the higher the responsiveness of demand, the more the firm stands to lose if it is caught with too low a price in the second period. Since costs are effectively indexed, a low nominal price can entail a loss which increases with the elasticity of demand. Caution should be exercised, however, in interpreting the relationship, because any firm can always pursue a two-price policy. High-elasticity firms are likely to exercise this option more readily than low-elasticity firms, both because the one-price policy is more risky for them and because they usually have lower adjustment costs. As the variability of the price level rises, one would generally expect many high-elasticity firms to switch from one- to two-price policies, but low-elasticity firms to continue pursuing one-price policies, albeit at higher prices.

---

[19] The crux of the proof is showing, through repeated use of Tchebychev's rearrangement inequality, that $R'(\alpha)>0$ in the neighborhood of 0. Once this relationship is established, it is sufficient to demonstrate through a continuity argument that $R(\alpha_2)>R(\alpha_1)$ for $\alpha_2>\alpha_1$.

[20] Recall that the real monopoly price is given by $bc/(b-1)$.

[21] This appears to be true not only in terms of percentage, but also in absolute terms.

## IV. Sensitivity of Propositions to Functional Forms in the Model

At this point it is useful to explain how the foregoing propositions depend on specific functional forms used in the analysis. Two assumptions require attention: constant price elasticity of demand and constant marginal cost.

The assumption that the price elasticity of demand is constant ensures that demand approaches infinity as price tends to zero, and as such is directly responsible for the asymmetry in the profit function between upward and downward displacements from the monopoly price. If elasticity were not constant but decreasing in price, the asymmetry would be even greater. If, however, elasticity were increasing, as is the case for the linear demand function, the asymmetry in the profit function could go either way.

The other functional form assumption, that real marginal cost is invariant to the scale of production, also plays a role in the observed asymmetry in the profit function. But if marginal cost were an increasing function of scale, the observed asymmetry in the profit function would be even greater. This is so because the opportunity cost from charging a price below the monopoly price would then exceed the opportunity cost from charging a price above the monopoly price to a greater extent than if marginal cost were constant. In contrast, if marginal cost were a decreasing function of scale, a counteracting asymmetry would be present. In this case, the asymmetry in the profit function could go either way.

In sum, the three propositions herein hold with any nonincreasing elasticity demand function and any nondecreasing marginal cost function. They do not necessarily hold, however, if over some range elasticity is increasing or marginal cost decreasing. In this case, the firm may exhibit deflationary bias, and its pattern of price rigidity may be fundamentally different.

## V. Summary

This paper has demonstrated that the prices of monopolistic firms with nonincreasing price elasticity of demand and nonincreasing marginal cost can be relatively more rigid downward than upward. It has also shown that such firms raise their prices to a greater extent when they expect inflation than they lower them when they expect equivalent deflation. Furthermore, it has established that they raise their prices in response to an increase in the degree of price level uncertainty, defined in terms of a geometric-mean-preserving spread. These findings are all attributable to the property that profit functions are asymmetric with respect to price, in the sense that the loss from charging a price below the monopoly price is greater than the loss from charging a price above it by a proportionally equivalent amount.

Of the three findings, the first two would be very difficult to test empirically if only because deflations almost never occur in modern economies. On the other hand, the last finding can, in principle, be tested, if one is able first to determine the particular sectors of the economy for which the model herein represents a reasonable abstraction. The available macroeconomic studies based on highly aggregated data indicate that the variability and rate of inflation are positively correlated,[22] and the third finding may point to one possible explanation for this empirical relationship. In periods when the time path of the price level is relatively uncertain, say, because of ongoing institutional changes or political instability, the asymmetry in some firms' profit functions would cause them to raise their prices, thereby putting upward pressure on the aggregate price level.

As stated earlier, however, other firms' adjustment policies might exhibit biases in opposite directions from those found in this paper. So it is impossible to say, at the present time, what cumulative effects the biases generated by asymmetries in firms' profit functions would have on the course of the general price level. But one can reasonably expect that, since conditions for perfect symmetry in profit functions are very stringent, the general price level would be influenced by individual firms, both directly through the actual prices they generate and indirectly through the effects of price adjustments on expectations. Given the vital importance of these effects for macroeconomic policy, it would be useful to examine the biases in the price-adjustment policies of various other types of business firms found in the economy.

---

[22] For some cross-country evidence, see, for example, Dennis Logue and Thomas Willett (1976).

In a perfectly competitive economy, where prices determined by impersonal market forces clear markets at each instant, biases of the kind discussed herein would not exist. Nor would they in an economy with a monopolistic sector composed of firms which can adjust their prices at no cost; firms would passively respond to changes in the price level, exerting no influence on it themselves. In actual economies, however, most firms set their own prices, often with the intention of keeping them fixed for substantial lengths of time. Here not only do price level movements affect individual price adjustments, but these adjustments influence the course of the price level.

## REFERENCES

**Cagan, Phillip,** *Persistent Inflation: Historical and Policy Essays*, New York: Columbia University Press, 1979.

**Gordon, Robert J.,** "Output Fluctuations and Gradual Price Adjustment," *Journal of Economic Literature*, June 1981, *19*, 493–530.

**Intriligator, Michael D. and Sheshinski, Eytan,** "Toward a Theory of Planning," mimeo., 1979.

**Logue, Dennis E. and Willett, Thomas D.,** "A Note on the Relation Between the Rate and Variability of Inflation," *Economica*, May 1976, *43*, 151–58.

**Means, Gardiner C.,** "Simultaneous Inflation and Unemployment: A Challenge to Theory and Policy," in his et al., *The Roots of Inflation*, New York: Burt Franklin, 1975, 1–31.

**Okun, Arthur M.,** *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

**Rothschild, Michael and Stiglitz, Joseph E.,** "Increasing Risk: I. A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

**Sheshinski, Eytan and Weiss, Yoram,** "Inflation and Costs of Price Adjustment," *Review of Economic Studies*, June 1977, *44*, 287–303.

**Solow, Robert M.,** "On Theories of Unemployment," *American Economic Review*, March 1980, *70*, 1–11.

**Stigler, George and Kindahl, James K.,** *The Behavior of Industrial Prices*, New York: National Bureau of Economic Research, 1970.

# Why Interest Rates Rise When an Unexpectedly Large Money Stock is Announced

*By* DONALD A. NICHOLS, DAVID H. SMALL, AND CHARLES E. WEBSTER, JR.*

It is widely accepted that the announcement of an unexpectedly large growth in the money supply will lead to an immediate increase in interest rates. This "announcement effect" is often taken as confirmation of the monetarist view that the way to bring down interest rates, even in the short run, is to reduce the rate of money growth. In this paper we provide an alternative explanation of the announcement effect that is consistent with the Keynesian view that, in the short run, the way to bring down interest rates is to expand the money supply. We do not address the issue of inflationary expectations in a formal way, though, in the long run, we would expect the well-known Fisher effect to hold.

Looking closely at the announcement effect, it is clear that whatever may be the true short-run relationship between money growth and interest rates, it cannot be inferred from a simple correlation of announcements of changes in the money stock with the resulting changes in interest rates. This is because of several properties of the announcement phenomenon that we capture in the model below. First, the money supply is not changed by the announcement, though interest rates are. The announcement provides new information, not new money. A complete analysis of the announcement effect must examine the link between information and interest rates, as well as that between money and interest rates. Since interest rates change with the announcements and the money supply does not, the new information in the announcement must cause money demand to change. Second, as pointed out in Nichols (1983), the announcement provides the same, possibly surprising, information about the quantity of money demanded as it does about the quantity supplied. The quantities supplied and demanded are known to be equal at the observed market interest rate before the announcement. If the announcement reveals a surprisingly large supply, it also reveals a surprisingly large demand. Because of this, we have used the neutral term money stock in the title of the paper to refer to a monetary aggregate, rather than the more common but misleading term, money supply.

We capture these essential characteristics of the announcement effect in a model that has the Keynesian properties of a fixed price level and a demand for money that is a decreasing function of the rate of return on securities. The rate of return includes a capital gains term that provides the link between present and future periods. Expectations of future variables are rational in the sense of John Muth (1961). That is, expectations of future variables are formed using the same model used to determine the equilibrium.

The important result we derive is that, *ceteris paribus*, the announcement of an unexpectedly large growth in the supply of money will cause current real interest rates to rise if the shock to money demand revealed by the announcement is expected to dissipate more slowly than the equally large shock to money supply that is simultaneously revealed. The reasoning behind this result is that an expected increase in the future real demand for money relative to supply causes expected future real interest rates to rise, which increases the demand for money in the present, and thus increases present interest rates as wealth holders try to avoid the losses they will realizes from holding securities when interest rates rise. When this result holds, it would be more informative to refer to the weekly data on the monetary aggregates as the money demand statistics rather than the money supply statistics.

---

*Professor of economics, University of Wisconsin; assistant professor of economics, University of Wisconsin and visiting assistant professor, Washington University-St. Louis; and assistant professor of economics, Washington University-St. Louis, respectively. This research was supported by grant no. 99-0-2289-50-11 from the National Commission for Employment Policy. We thank an anonymous referee for helpful suggestions.

Our result provides a plausible explanation of the announcement effect. These effects became common only in the late 1970's which is when the Federal Reserve began following a fixed money growth rule. When the supply of money is expected to follow a fixed rule, deviations from the path implied by the rule can be expected to be temporary. There is less reason to expect the deviations in money demand to be temporary, however, hence it is rational to expect future interest rates to rise when a large money shock is revealed. Note, however, that an increase in the Federal Reserve's money growth target would still bring down interest rates in this Keynesian model, since such a change would increase the expected supply of money by more then the expected demand.

In Sections I and II, we derive the implications of unexpected monetary announcements on the current and expected future interest rates using a Keynesian model. In Section III, we contrast our results with those of a simple monetarist model.

## I. The Effect of an Announcement on the Current Real Interest Rate

The problem is modeled in discrete time. The demand for money is assumed to be a decreasing function of the expected return on competing assets, which we will call bonds. The expected rate of return on bonds equals its one-period interest rate plus the expected capital gain which will exist if the next period's interest rate is expected to differ from this period's rate. For convenience, the interest rates in the capital gains term are expressed in linear form. The money demand function is

$$(1)\qquad M_t^d = \alpha r_t + \beta(r_t - E_t r_{t+1}) + d_t,$$

where $M_t^d$ = the quantity of money demanded at time $t$; $r_t$ = the market interest rate available at time $t$ on bonds held from time $t$ to $t+1$; $E_t r_{t+1} = E[r_{t+1}|I_t]$: the mathematical expectation of $r_{t+1}$ conditional on information available at time $t$; $d_t$ = shock to money demand; and $\alpha < 0$, $\beta < 0$. The information set is assumed to include the values of all variables from earlier periods as well as $r_t$. The inclusion of $r_t$ captures the continuous observation of the interest rate by financial market participants.

We assume that the money demand shock follows a first-order autoregressive process that is known by all market participants:

$$(2)\qquad d_t = \gamma d_{t-1} + u_t$$

where $0 < \gamma < 1$ and $u_t = a$ white noise random variable with zero mean and finite variance.

The deterministic component of the supply of money ($\overline{M}$) can be assumed to be determined by the Federal Reserve's target. The supply of money also has a random component that follows a first-order autoregressive process like that of the demand for money. The parameter $\delta$ is the speed with which the Federal Reserve is known to offset any accumulated shocks from the past.

$$(3)\qquad M_t^s = \overline{M}_t + s_t;$$

$$(4)\qquad s_t = \delta s_{t-1} + v_t,$$

where $0 < \delta < 1$ and $v_t$ is a white noise random variable with zero mean and finite variance.

The condition for equilibrium in the money market is

$$(5)\quad \overline{M}_t + s_t = \alpha r_t + \beta(r_t - E_t r_{t+1}) + d_t.$$

To solve this relation for $r_t$, we must first develop an expression for $E_t r_{t+1}$. Note that if $\beta = 0$, then $r_t$ depends only on current variables and not on expectations of future variables as is the case when there is no announcement effect.

Define the operator $B$ such that

$$(6)\qquad B^{-j} E_t X_s = E_t X_{s+j}.$$

Then rearrange (5), define $\lambda = (\alpha + \beta)/\beta$, take expected values of both sides conditional on an information set that includes $\overline{M}_t$

(but not $d_t$ or $s_t$) to obtain

$$(7)\quad (1-\lambda B)E_t r_{t+1} = (\overline{M}_t + E_t s_t - E_t d_t)/\beta.$$

In the Appendix, the general solution for $E_t r_{t+1}$ is presented. For notational simplicity, equation (8) is the solution after we have imposed $E_t v_{t+i} = E_t u_{t+i} = 0$ for positive $i$ and $\overline{M}_{t+i} = \overline{M}$ for all $i$.

$$(8)\quad E_t r_{t+1} = \frac{1}{\beta}\left[\frac{\delta E_t s_t}{\lambda-\delta} - \frac{\gamma E_t d_t}{\lambda-\gamma} + \frac{\overline{M}}{\lambda-1}\right].$$

Note that the higher the deterministic component of the future money supply $\overline{M}$, the lower the future interest rates are expected to be. Also, the current shocks have no effect on expected future interest rates unless the shocks persist, that is, unless $\delta \neq 0$ or $\gamma \neq 0$.

The expression for the current interest rate can be obtained by substituting (8) into (5):

$$(9)\quad r_t = \frac{1}{\beta\lambda}\left[\frac{\lambda\overline{M}}{\lambda-1} + s_t - d_t - \frac{\gamma E_t d_t}{\lambda-\gamma} + \frac{\delta E_t s_t}{\lambda-\delta}\right].$$

Note that this expression contains the actual shocks to the money supply and demand functions as well as the expected values of those shocks. Before the money demand statistics are announced, their expected levels have an influence on the market interest rate. So do the actual (though unknown) levels. Thus both sets of variables must appear in the solution. Note also that the effect of the known money supply ($\overline{M}$) on the interest rate is negative since $\beta < 0$.

Before deriving the effect on the current interest rate of the announcement of unexpected money growth, we note that any underestimation of the money supply shock $s$ is equal in magnitude to the underestimation of the money demand shock $d$. While this may be intuitively clear, it can be derived from equation (9) as well. Taking expectations of both sides of (9) and subtracting this result from (9) yields (10). Note in taking expectations of both sides of (9) $E_t[r_t] = r_t$, since $r_t$ is an element of $I_t$.

$$(10)\quad \begin{aligned} s_t - E_t s_t &= d_t - E_t d_t \\ v_t - E_t v_t &= u_t - E_t u_t. \end{aligned}$$

Upon announcement, the change in the expectation of the money supply, $s_t - E_t s_t \equiv dE_t s_t$, must equal the change in the expectation of money demand, $d_t - E_t d_t \equiv dE_t d_t$. Thus the effect of the revelation that the expectations were wrong on the interest rate is given in (11). (It should be noted that we are ignoring the reporting lag of the Federal Reserve. On Friday, the Fed reports the money supply for the week ending on the previous Wednesday. This delay does not affect our qualitative results.)

$$(11)\quad \frac{dr_t}{dE_t s_t} = \frac{1}{\beta}\left[\frac{\delta}{\lambda-\delta} - \frac{\gamma}{\lambda-\gamma}\right].$$

Since $\lambda > 1$, $\gamma < 1$, $\delta < 1$, $\beta < 0$, the sign of this expression is the same as the sign of $\gamma - \delta$. In words, if the shock to money demand is expected to persist longer than the shock to money supply, interest rates will rise when an unexpectedly large increase in the money stock is announced. This is the Keynesian explanation of the announcement effect.

Because the Federal Reserve now follows a policy of controlling the money stock, investors might expect that the shock to money supply would be more transitory than that to money demand. Figure 1 shows the extreme case of this where $\delta$ is close to zero and $\gamma$ is close to one. In this case, investors view the supply shock as completely transitory and the demand shock as almost permanent. Before the announcement, investors observe the market-clearing interest rate of $r_0$. They expect a figure of $\overline{M}$ to be announced as the prevailing money supply. In fact, $M_t^s$ is then announced, revealing an error of $v_t$ in their prior expectations. Both demand and supply are revealed to be larger by $v_t$ than anticipated.

If the shock to demand is viewed as mostly permanent, next period's money demand is now expected to be close to $M_t^d$ as drawn (except for the effect of the capital gains term). If the shock to supply is viewed as mostly transitory, next period's supply is expected to be close to $\overline{M}$ yielding an interest rate of $r_1$ next period. Because $r_1$ is larger than $r_0$, losses will be sustained by anyone holding interest bearing securities from now

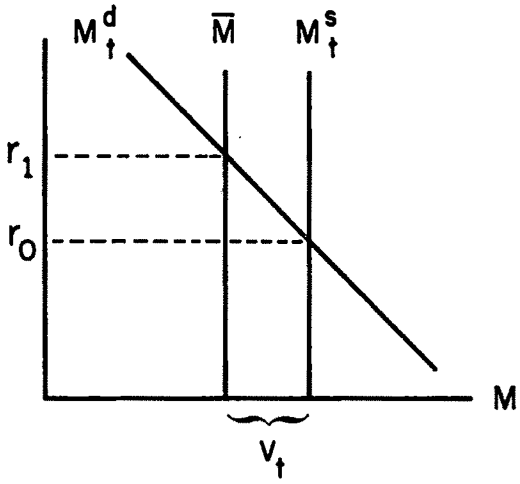


FIGURE 1

to next period. The attempt to avoid these losses causes demand to increase (unshown) and the current rate to rise. If the time periods are short, $r_0$ can rise to a level close to $r_1$ upon the announcement.

## II. The Effect on Expected Future Interest Rates

If the shocks are thought to be temporary ($\gamma < 1$; $\delta < 1$), the announcement effect on expected future real interest rates is weaker, the farther in the future we look. This result depends, of course, on the specification of the supply and demand shocks.

As shown in the Appendix, the effect of the announcement on $E_t r_{t+j}$ is

$$(12)\quad \left.\frac{dE_t r_{t+j}}{dE_t v_t}\right|_{dE_t v_t = dE_t u_t} = \frac{1}{\beta}\left[\frac{\delta^j}{\lambda-\delta} - \frac{\gamma^j}{\lambda-\gamma}\right].$$

Thus the signs of these terms for all future interest rates depend, as before, on the relative size of $\delta$ and $\gamma$. If $\gamma > \delta$, the expected future real interest rates all rise if the announced money stock is larger than expected, since $\beta < 0$. This holds regardless of any expected capital gains or losses because, with $\gamma > \delta$, a larger than expected money stock announcement is a signal that there will be excess demand for money in the future unless future interest rates change from their *ex ante* expected values. Note also that as $j$ rises, the expression gets smaller. Differentiating (12) with respect to $j$ yields

$$(13)\quad \frac{1}{\beta}\left[(log\,\delta)\frac{\delta^j}{\lambda-\delta} - (log\,\gamma)\frac{\gamma^j}{\lambda-\gamma}\right] < 0 \qquad \text{if } \gamma > \delta.$$

Thus the effect on future interest rates is smaller, the further into the future we look.

One cautionary note for the interpretation of these results is that the prediction of the effect of an announcement on the current rate $r_t$ is smaller than its effect on $r_{t+1}$. The result that the effect on $r_{t+j}$ be greater than on $r_{t+j+1}$ holds for all $j \geq 1$, but not for $j = 0$. This imposes an awkward qualification to the general result that the effect should be weaker on interest rates of longer maturities. Our discrete model provides no natural interpretation of the length of a time period. If we thought of a continuous time model as the limiting case of our model when the length of a time unit approached zero, then the model would predict an instantaneous jump of all rates—including the overnight rate—if a surprisingly large rate of money growth was announced. For time units of longer length, the overnight rate might rise less than the weekly rate.

The economics of the model tells us that the rates rise because the Federal Reserve is expected to offset the effects of the money shock faster than the money demand shock is expected to dissipate. If the pattern of evaporation of the supply and demand shocks differs from the exponential decay pattern we have used, the analysis might predict a larger effect on rates with a maturity other than the shortest, and any empirical test would have to permit a rather general specification of the stochastic nature of the shocks. If, for example, the quickest Federal Reserve reaction is expected in two days, it is the two-day interest rate that should be affected most by the announcement.

## III. Elements of a Test

We have not developed a conclusive formal test to determine whether our explanation of the announcement effect is of greater empirical importance than the monetarist explanation. But in this section we point out some characteristics that are of use in constructing such a test.

First, we note that our explanation of the announcement effect is that an increase in the money stock when the authorities are following a strict money growth rule is a signal that the expected future demand for money has increased, but the supply has not. If these conditions are assumed to hold in a monetarist model with flexible prices, the announcement would provide a signal that future prices would fall, not rise, since an increase in the demand for money is a contractionary, not an expansionary, force. In this case, the inflationary premium in future interest rates should fall, not rise. Thus the prediction of our model is opposite to that of the monetarist model under the plausible conditions that were stated. This effect would be more readily apparent if the weekly money stock figures were referred to as the money demand statistics, rather than the money supply statistics.

One qualification to this result is that our model is based on the assumption that the price level is known when the money stock is announced. If, in a monetarist model, this assumption is dropped, things get more complicated. It is for this reason we do not claim that our proposed test is conclusive.

Monetarists have argued that an unexpectedly large money stock announcement is a signal that it is the future money supply, not demand, that will be increased. This possibility is admitted. We note, however, that it should be of lesser importance when the Federal Reserve is trying to follow a fixed growth rule than when it is not. Robert Lucas (1976) suggests that the way to test macroeconomic hypotheses is to examine them in two different policy regimes. If recent history could be divided into two periods (possibly divided in October 1979), in one of which a money growth rule was pursued more avidly than in the other, then the predictions of the two competing hypotheses would be in direct opposition. The monetarist hypothesis would predict a smaller announcement effect when a money growth rule is followed and the Keynesian hypothesis would predict a larger effect.

While casual empiricism would appear to support the Keynesian hypothesis on this, we caution that we have not conducted or even specified a formal test. Our objectives have been more limited, namely, to provide a possible counterexample to the claim that the existence of the announcement effect confirms the monetarist position that an increase in money growth rates must raise short-term interest rates.

## IV. Conclusion

A common interpretation of the announcement effect—the fact that interest rates rise when unexpectedly large money stocks are announced—is the monetarist one that an increase in the money growth rate raises the expectations of inflation that are part of nominal interest rates. A competing explanation offered here is consistent with the Keynesian claim that interest rates fall when the money supply is expanded. The competing explanation holds in periods when the central bank is following and is expected to follow a fixed money growth rule. In this case, the money stock announcements convey information about unexpected strength in current and future money demand. This will increase expected future interest rates and their increase will pull up the current interest rate as investors try to avoid capital losses.

Thus the existence of an announcement effect should not be interpreted as it is has been by Jacob Grossman (1981) that it is conclusive evidence that "The long lag in the adjustment of [interest] rates to money expansion as suggested in Friedman and Schwartz [1963] and Cagan [1972] has been considerably shortened" (p. 421, fn. 33). The announcement effect concerns the effect of changes in *information* on interest rates, not of changes in the existing *money stock*. If the announcement effect is more pronounced today than in years past, it is probably because the Federal Reserve is now expected to follow more closely a constant growth rate rule. A surprisingly large money stock announcement is then interpreted as a sign that money is about to be tightened.

## APPENDIX

Detailed derivations of the following results are available from the authors upon request. We have used the method of linear operators as presented by Thomas Sargent (1979, chs. 9 and 14). Alternatively one can employ the method of successive substitution.

The general form of equation (8) is

$$\text{(A1)}\quad E_t r_{t+1} = \frac{1}{\beta\lambda}\sum_{i=0}^{\infty}\left(\frac{1}{\lambda}\right)^i M_{t+1+i} + \frac{1}{(\lambda-\delta)\lambda}\sum_{i=0}^{\infty}\left(\frac{1}{\lambda}\right)^i E_t v_{t+1+i} + \delta E_t s_t - \frac{1}{(\lambda-\gamma)\lambda} \times \sum_{i=0}^{\infty}\left(\frac{1}{\lambda}\right)^i E_t u_{t+1+i} + \gamma E_t d_t.$$

This can be derived by dividing (7) by $1-\lambda B$ and taking the "forward" expansion since $\lambda > 1$. In the derivations one uses the equalities

$$\text{(A2)}\quad \frac{E_t s_t}{1-\lambda B} = \frac{1}{1-\lambda B}E_t\left(\frac{v_t}{1-\delta L}\right) = \frac{1}{\lambda-\delta}\left[\frac{\lambda E_t v_t}{1-\lambda B} - E_t\left(\frac{\delta}{1-\delta L}v_t\right)\right],$$

where $L$ is the lag operator. In doing "backward" expansions with $\delta < 1$, the operator $B$ can be substituted for the operator $L$ since past values of $s_t$ are known.

Equation (12) of the text can be derived from (A1) by first substituting

$$s_t = \frac{v_t}{1-\delta L} = \sum_{i=0}^{\infty}\delta^i v_{t-i}$$

and

$$d_t = \frac{u_t}{1-\gamma L} = \sum_{i=0}^{\infty}\gamma^i u_{t-i}$$

into (A1), multiplying both sides by $B^{-j+1}$ and differentiating with respect to $E_t v_t$ and $E_t u_t$ while imposing $dE_t u_t = dE_t v_t$.

## REFERENCES

**Cagan, Phillip,** *Channels of Monetary Effects on Interest Rates*, New York: Columbia University Press, 1972.

**Conrad, William E., Jr.,** "Initial Treasury Bill Market Response to Money Stock Announcement," processed, Board of Governors of the Federal Reserve System, 1978.

**Friedman, Milton and Schwartz, Anna J.,** "Money and Business Cycles," *Review of Economics and Statistics*, February 1963, Suppl., *45*, 32–64.

**Grossman, Jacob,** "The 'Rationality' of Money Supply Expectations and the Short-Run Response of Interest Rates to Monetary Surprises," *Journal of Money, Credit, and Banking*, November 1981, *13*, 409–24.

**Lucas, Robert E. Jr.,** "Econometric Policy Evaluation: A Critique," in Karl Brunner and Allan H. Meltzer, eds., in *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

**Muth, John F.,** "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

**Nichols, Donald,** *Active Economic Policy*, Belmont: Wadsworth Publishing, 1983.

**Sargent, Thomas J.,** *Macroeconomic Theory*, New York: Academic Press, 1979.

**Sivesind, Charles M.,** "Efficient Markets and the Reaction of Securities, Markets to the Weekly *M*1 Announcement," processed, Federal Reserve Bank of New York, 1978.

# Labor Market Contacts, Unanticipated Wages, and Employment Growth

*By* RONALD S. WARREN, JR.*

There are two alternative, but not mutually exclusive, explanations for observed fluctuations in aggregate employment. The Keynesian, disequilibrium theory emphasizes the importance of quantity constraints, job rationing, and the resulting queues of involuntarily unemployed suppliers of labor that arise from exogenous decreases in investment, aggregate demand, output, and the demand for labor. The new classical, equilibrium or search theory focuses on the optimizing responses of suppliers of labor to deviations of actual from anticipated wage offers that are generated from innovations in nominal aggregate demand.[1] The choice between (or relative importance of) these two explanations is of considerable importance not only for the proper interpretation of cyclical phenomena in the labor market, but also for the design of macroeconomic policy. For example, proponents of the disequilibrium viewpoint stress the importance of activist, job creation intervention by the government when aggregate demand falls short of full-employment output. On the other hand, adherents to the search-theoretic perspective typically advocate predictable monetary growth and automatic fiscal stabilizers which are intended to minimize the (temporary) misinformation provided by unanticipated changes in nominal wages and prices.[2]

In the past, there have been few tests of these two hypotheses and the empirical results have been mixed. Moreover, previous investigations have attempted primarily to explain features of the labor market—like the quit rate or the duration of unemployment—which are only of intermediate concern to the central task of accounting for the time paths of employment and unemployment. For example, Donald Parsons (1973) presented evidence from twenty-seven manufacturing industries that the quit rate was significantly related to the vacancy rate (probability of receiving a job offer), but not affected by misperceptions about equilibrium wages. Similarly, John Barron (1975), using aggregate data from U.S. manufacturing, found that the average duration of unemployment was inversely related to the number of vacancies and inferred from these results that unanticipated changes in the wage offer distribution were unimportant.[3] More recently, Anders Björklund and Bertil Holmlund (1981) provided some support for the proposition that both job opportunities and wage misperceptions significantly affected the duration of unemployment in U.S.

---

*Assistant professor of economics, University of Virginia. I thank Daniel Hamermesh, William Johnson, John Seater, Robert Solow, and two anonymous referees for helpful comments on an earlier version of this paper. John Carlson kindly provided data on one of the wage control variables. Of course, responsibility for remaining errors is mine.

[1] There are actually three distinct representations of the search theory. The intertemporal substitution hypothesis advanced by Robert Lucas and Leonard Rapping (1970) states that current labor supply is a function of the difference between current and discounted, expected *future* wages. Alternatively, the wage misperception theory contends that current labor market behavior depends upon the difference between actual and expected *current* wages. Within this genre, Milton Friedman (1968) and Lucas (1973) emphasize the misperception of *real* wages while Dale Mortensen (1970) stresses the deviation of actual from anticipated *nominal* wages. This paper addresses the real wage misperception theory rather than the intertemporal substitution hypothesis. John Seater (1982) has recently provided preliminary tests of all three versions of the neoclassical approach within one estimating equation.

[2] That is, the new classical approach examines how stabilization policy rules might be designed to *avoid* cycle-generating misperceptions, while the Keynesian framework focuses on how discretionary policy should respond to *existing* misperceptions.

[3] However, Anthony Santomero and Seater (1978, pp. 521–22) have criticized the test procedures used by Parsons and Barron.

manufacturing, when expected wages were modelled as an autoregressive-moving average (ARMA) process.

The purpose of this paper is to examine the validity of these two perspectives in the context of a modified version of a simple model of employment dynamics. Aside from parsimony, the virtues of this model for the issue at hand include its familiarity (to students of aggregate models of the labor market) and the facility with which its parameters can be related to concepts like the labor turnover rate, efficiency of job search, and the equilibrium or "natural" unemployment rate. Section I sets out the theoretical model in which employment growth is determined by the turnover rate, job search efficiency, and the volume of contacts between searching firms and workers. Deviations of actual from expected wages affect employment indirectly by influencing the quit rate and the proportion of contacts which result in accessions. Section II discusses data sources and estimation technique, and presents the empirical results.

## I. The Model

The basic theoretical framework for this study is provided by the well-known model of labor market dynamics developed by Richard Lipsey (1960), Charles Holt and Martin David (1966), Edmund Phelps (1970), Dale Mortensen (1970), and others. There are four endogenous variables—the level of unemployment ($U$), the level of unfilled vacancies ($V$), the level of employment ($E$), and the change in employment ($\Delta E$)—which are determined by three identities and one difference equation. There are two exogenous variables—the total number of jobs ($J$) and the labor force ($L$). The three identities in the model are

$$V \equiv J - E, \tag{1}$$

$$U \equiv L - E, \tag{2}$$

$$E \equiv \Delta E + E_{-1}, \tag{3}$$

where $E_{-1}$ is the level of employment in the previous period.

Each period, new entrants and reentrants to the labor force and formerly employed workers who quit or were laid off conduct a random search of firms which are seeking to fill job vacancies. This search process takes place under conditions of imperfect information about the location of vacancies and the true probability distribution which generates observed values of wage offers for labor of a given type. The setting in which an employer with a vacancy and a qualified worker are matched in this model is essentially an accidental meeting.[4] Thus, it is distinct from the neoclassical set up in which an omniscient auctioneer assists searching firms and persons in defining the terms of a labor contract of positive but finite duration. Because each participant in this matching process is imperfectly informed about the (equilibrium) characteristics of his optimal match, all vacancies are not instantaneously filled and all available workers are not immediately employed. Firms and workers weigh the cost of finding another match against the discounted expected value of negotiating a contract in the future on more favorable terms.

The steady-state stocks of unfilled vacancies and unemployed workers are those that are consistent with an equality between the flows of accessions to and separations from employment. Accessions (new hires and rehires) depend upon the number of contacts per period between firms and potential workers and the rate (denoted $\beta$) at which these contacts result in new labor contracts. It is assumed that the number of contacts made each period is $(UV)^{1/2}$ so that accessions equal $\beta(UV)^{1/2}$. Separations are composed of quits and layoffs, which are inversely related over the business cycle. If the sum of quits and layoffs is assumed to be proportional to the level of employment, then the number of separations is given by $\gamma E$, where $\gamma$ is the turnover rate.[5]

The net change in employment ($\Delta E$) is simply the difference between accessions to and separations from employment:

---

[4] This characterization was put forward by Mortensen (1976).

[5] The assumptions concerning accessions and separations were made by Lipsey (p. 15) and C. Duncan MacRae and Stuart Schweitzer (1970, p. 89).

$$(4)\qquad \Delta E = \beta (UV)^{1/2}_{-1} - \gamma E_{-1},$$

or

$$(5)\qquad \Delta E/E_{-1} = \beta (U/E \cdot V/E)^{1/2}_{-1} - \gamma .$$

The steady-state or equilibrium unemployment rate $u^* = (U/L)^*$ can be solved for by setting $\Delta E/E_{-1} = 0$:

$$(6)\qquad u^* = (\gamma/\beta)^2 E^2/VL.$$

Thus, holding employment, job vacancies, and the labor force constant, the equilibrium unemployment rate is positively related to the turnover rate and negatively related to the efficiency of job search.

As presently formulated, there are (at least) two related shortcomings of this model. First, employment growth and equilibrium unemployment are determined by the exogenous flows of *quantities* of jobs and labor force participants, the turnover rate, and search efficiency. In particular, relative or unanticipated *wages* play no role in the observed operating characteristics of the labor market, contrary to neoclassical theory. Second, the turnover rate ($\gamma$) and search efficiency ($\beta$) are specified as *parameters* when it is more reasonable to view them as *functions* of wages and government policies. For example, there is a corpus of theory and considerable microeconomic evidence which suggest that quits and layoffs are determined in part by the difference between actual and potential (human capital) wages.[6] Moreover, layoffs—especially among youth and other unskilled workers—may be affected by changes in the minimum wage or other policy instruments. Similarly, search efficiency—the proportion of contacts per period that results in employment—may also be influenced by actual relative to expected wages, as well as government policies (like unemployment insurance) which reduce the cost of being unemployed.

Taken together, these two criticisms suggest a way in which both quantity constraints and wage misperceptions can simultaneously affect employment growth and unemployment in the present model. If it can be assumed that unanticipated wages affect turnover and search efficiency in a linear fashion, we can write, respectively:

$$(7)\qquad \gamma = \gamma_0 + \gamma_1 (w - w^e)_{-1},$$

$$(8)\qquad \beta = \beta_0 + \beta_1 (w - w^e)_{-1},$$

where $w$ is the observed real wage, $w^e$ is the expected real wage, and $\gamma_0$, $\gamma_1$, $\beta_0$, and $\beta_1$ are parameters.

The argument that unexpectedly high wages reduce the quit rate and, hence, labor turnover implies $\gamma_1 < 0$. Similarly, if current wages or wage offers are higher than anticipated, this will, according to search theory, increase the proportion of worker-firm contacts which results in acceptances of offers so that $\beta_1 > 0$.[7] These qualitative effects on labor turnover and search efficiency imply that real wages which are higher than expected result in a (temporary) decrease in *observed* unemployment. If, in the long run, wage expectations are rational in the sense that $w = w^e$, then *equilibrium* unemployment is given by (6) and cannot be reduced systematically by aggregate demand policies that affect the distribution of nominal wages.

Substituting (7) and (8) into (5), we have

$$(9)\qquad \Delta E/E_{-1} = [\beta_0 + \beta_1 (w - w^e)_{-1}] X_{-1} - [\gamma_0 + \gamma_1 (w - w^e)_{-1}],$$

$$(10)\qquad \Delta E/E_{-1} = \pi_0 + \pi_1 X_{-1} + \pi_2 (w - w^e)_{-1} + \pi_3 (w - w^e)_{-1} X_{-1},$$

where $X_{-1} \equiv (UV/E^2)^{1/2}_{-1}$, $\pi_0 = -\gamma_0$, $\pi_1 = \beta_0$, $\pi_2 = -\gamma_1$, and $\pi_3 = \beta_1$. The qualitative implications of the wage misperceptions hypothesis for the coefficients in (10) are that $\pi_2$ and $\pi_3 > 0$. In addition, since $\gamma_0$ and $\beta_0$ are expected to be positive, we expect $\pi_0 < 0$ and $\pi_1 > 0$. Therefore, estimates of the coefficients in (10) and their associated standard errors can be used to test the validity of the quantity constraint and wage misperceptions theories of employment dynamics.

---

[6] For a review of this literature, see Parsons (1977).

[7] Since (5) should be interpreted as a reduced form, the prediction that $\beta_1 > 0$ depends on the maintained hypothesis that labor supply is more responsive to real wage misperceptions than is labor demand, as Robert Barro (1980, p. 1404) has noted in a slightly different context. In particular, when the real wage is unexpectedly high, the quantity demanded of labor should be low, resulting in fewer offers of employment as a proportion of labor market contacts.

## II. Empirical Results

To estimate equation (10), data on employment, unemployment, job vacancies, wages, and expected wages were required. Although economywide, monthly data on employment, unemployment, and wages were, of course, readily available over a long time span, information on job vacancies was not. However, from April 1969 to December 1973, the U.S. Department of Labor, Bureau of Labor Statistics (BLS) published monthly estimates of job vacancies in manufacturing industries. Moreover, these data have been used previously by Barron and Björklund-Holmlund to assess the relative impact of job opportunities and wage misperceptions on the duration of unemployment and by William Goodwin and John Carlson (1981) to evaluate the effects of wage-price controls. Thus, although data limitations restrict considerably the time period and sector to which the present investigation can be applied and the extent to which seasonal and cyclical effects can be controlled, they facilitate comparisons with previous research on closely related issues. Data on employment and unemployment were also obtained from BLS sources. The wage variable used was average hourly earnings in manufacturing, adjusted for overtime hours, divided by the Consumer Price Index. All of these variables are *not* seasonally adjusted and are described more fully in Table 1.

TABLE 1—MEANS AND DEFINITIONS OF VARIABLES AND SOURCES OF DATA

| Variable | Mean | Definition and Source |
|---|---|---|
| $E$ | 19414.0 | Employment on manufacturing payrolls (in thousands); BLS, unpublished tabulations. |
| $U$ | 1099.6 | Number of persons unemployed whose last job was in manufacturing (in thousands); BLS, *Employment and Earnings*, various issues. |
| $V$ | 150.4 | Number of unfilled job vacancies in manufacturing (in thousands); BLS, *Employment and Earnings*, Vol. 20, Table E-1, p. 123. |
| $w$ | 2.86 | Hourly earnings excluding overtime in manufacturing divided by the Consumer Price Index for all-urban workers; BLS, as compiled by Data Resources, Inc., USCEN data bank. |
| $\hat{w}$ | 2.86 | Forecast of hourly earnings excluding overtime in manufacturing, based on a first-order, autoregressive, seasonal moving average model applied to the seasonal first differences of $w$. |

In order to test the hypothesis that unanticipated wages affect employment adjustments, a proxy variable for expected wages had to be constructed. To accomplish this task, wage predictions were generated from a seasonal ARMA process which was identified and estimated initially from 1960:1 to 1969:3, using monthly data on real average hourly earnings in manufacturing adjusted for overtime. Specifically, an ARMA (1,1) model with a seasonal MA(1) term was applied to the seasonal first differences of the wage variable.[8] A prediction from this estimated model was generated initially for 1969:4. To obtain successive predictions, the model was updated at monthly intervals

---

[8] That is, the forecasting model was $y - \phi y_{-1} = \varepsilon - \theta_1 \varepsilon_{-1} - \theta_{12}\varepsilon_{12}$ where $y \equiv (1 - B^{12})(1 - B)w$, $B$ is the backshift operator, and $\varepsilon_t$ is a white noise error process. It should be mentioned that a series on unanticipated *nominal* wages was also generated, using the time-series process and updating technique employed by Björklund and Holmlund (p. 124). The use of these data (which amounts to a test of the Mortensen 1970 variant of the misperceptions hypothesis) rather than unanticipated *real* wages did not affect the qualitative results reported below.

TABLE 2—EMPIRICAL RESULTS[a]

| Coefficients | I | II | III |
|---|---|---|---|
| $\pi - 0 = -\gamma_0$ | $-0.0589^b$ | $-0.0593^b$ | $-0.0731^b$ |
| | (2.8131) | (2.0316) | (3.5154) |
| $\pi_1 = \beta_0$ | $2.9744^b$ | $2.8009^b$ | $3.3879^b$ |
| | (2.8414) | (1.9364) | (3.3439) |
| $\pi_2 = -\gamma_1$ | $-0.6635$ | $-1.2073$ | $-1.1203$ |
| | (0.5729) | (1.0983) | (0.9996) |
| $\pi_3 = \beta_1$ | 35.9220 | 64.5110 | 60.2000 |
| | (0.6214) | (1.1742) | (1.0739) |
| $\pi_4 = -\gamma_2$ | | $-0.0122$ | |
| | | (0.3130) | |
| $\pi_5 = \beta_2$ | | 1.0080 | |
| | | (0.5170) | |
| $\pi_6 = -\gamma_3$ | | | $0.0002^c$ |
| | | | (2.4648) |
| $R^2$ | 0.14 | 0.27 | 0.23 |
| $F$ | $2.72^b$ | $3.73^b$ | $3.76^b$ |
| $D.W.$ | 2.28 | 2.58 | 2.49 |

[a] Absolute values of estimated $t$-statistics are shown in parentheses.

[b] Significantly different from zero at the 0.05 percent level, one-tailed test.

[c] Significantly different from zero at the 0.05 percent level, two-tailed test.

through 1973:11. Thus, for example, the prediction for the final period (1973:12) was based upon the ARMA (1,1) model estimated with data over the period 1960:1 through 1973:11.

Ordinary least squares (*OLS*) estimates of the model specified in equation (10) are reported in column I of Table 2. The goodness-of-fit of the model, as measured by the $R^2$ statistic, is acceptable in view of the percentage form of the dependent variable and the use of monthly observations. In addition, the overall $F$-statistic is significantly different from zero. Moreover, the $D.W.$ statistic does not indicate the presence of serially correlated residuals.

The estimated intercept ($\pi_0 = -\gamma_0$) is negative as expected, and significantly different from zero. That is, other things equal, an increase in the labor turnover rate would reduce employment growth and (by equation (6)) increase the equilibrium unemployment rate. The estimated coefficient ($\pi_1 = \beta_0$) on $X_{-1} \equiv (UV/E^2)^{1/2}_{-1}$ is positive, as the theory predicts, and also significantly different from zero. Thus, an increase in the volume of contacts by participants significantly increases employment growth and by implication decreases the equilibrium unemployment rate. The estimated coefficient ($\pi_2 = -\gamma_1$) on the unanticipated wage variable ($w - w^e$) (which shifts the turnover parameter) has an incorrect (negative) sign, but is not significantly different from zero. Moreover, the coefficient ($\pi_3 = \beta_1$) which affects the search efficiency parameter is correctly signed, but also insignificant. Thus, in contradiction to the predictions of the wage misperception hypothesis, this evidence suggests that deviations of actual from anticipated wages significantly affect neither the turnover rate nor the efficiency of job search. Hence, the time paths of employment growth and observed unemployment appear to be invariant with respect to mistakes in predicting real wages.

These estimates can be used to calculate the equilibrium or natural rate of unemployment in manufacturing over the sample period. In the context of the present model, equilibrium requires that $\Delta E/E_{-1} = 0$ *and* $w = w^e$ so that there are no wage misperceptions. From (10), therefore, average equilibrium unemployment $\bar{u}^* \equiv (\bar{U}/\bar{L})^*$ is given by

$$\bar{u}^* = (\hat{\gamma}/\hat{\beta})\bar{E}^2/\bar{V}\bar{L} \tag{11}$$

where bars over the variables denote sample means, and "hats" over parameters denote sample estimates. For the sample at hand, $\bar{u}^* = 0.0479$ so that the equilibrium unemployment rate in the manufacturing sector for the 1969:4–1973:12 period was approximately 4.8 percent. This estimate of the *natural* rate should be compared with the average *measured* rate in manufacturing of 5.7 percent over this sample period, which encompasses one complete business cycle. Thus, average "demand-deficient" unemployment in manufacturing was approximately 1 percent between 1969 and 1973. In addition, this estimate can be compared with Phillip Cagan's (1977) estimate of a 5.0 percent "full-employment" unemployment rate for the overall economy in 1973 and Robert

Barro's (1977) estimate of an economywide natural rate of 5.8 percent for the 1969–73 period.[9]

One potentially important characteristic of approximately half of the sample period under investigation was the existence of wage-price controls. The presence of controls might be expected to narrow the difference between actual and expected wages, thus biasing the previously reported test against the misperceptions hypothesis.[10] Moreover, Goodwin and Carlson have suggested that controls could increase the duration of job search and induce greater labor turnover by (incorrectly) increasing job seekers' perceptions of the number of vacancies. To test first the proposition that the controls reduced wage misperceptions, $w-w^e$ was regressed against a dummy variable ($D$) equal to one during the period covered by controls (1971:8 to 1973:12) and equal to zero otherwise. Although the estimated coefficient on the dummy variable was negative, it was not significantly different from zero ($t=0.6189$).[11] To test the Goodwin-Carlson hypothesis, it was assumed that controls would affect turnover ($\gamma$) and search efficiency ($\beta$) in the following manner:

$$(12)\quad \gamma=\gamma_0+\gamma_1(w-w^e)_{-1}+\gamma_2 D;$$

$$(13)\quad \beta=\beta_0+\beta_1(w-w^e)_{-1}+\beta_2 D.$$

Substituting (12) and (13) into (5), we have

$$(14)\quad \Delta E/E_{-1}=[\beta_0+\beta_1(w-w^e)_{-1}+\beta_2 D]X_{-1}-[\gamma_0+\gamma_1(w-w^e)_{-1}+\gamma_2 D],$$

$$(15)\quad \Delta E/E_{-1}=\pi_0+\pi_1 X_{-1}+\pi_2(w-w^e)_{-1}+\pi_3(w-w^e)_{-1}X_{-1}+\pi_4 D+\pi_5 DX_{-1},$$

where $\pi_4=-\gamma_2$ and $\pi_5=\beta_2$. Since the Goodwin-Carlson hypotheses imply that $\gamma_2>0$ and $\beta_2<0$, respectively, the qualitative expectations for the associated coefficients in (15) are $\pi_4<0$ and $\pi_5<0$.

Column II, Table 2, reports the results of estimating (15). Once again, the theoretical parameters $\pi_0=-\gamma_0$ and $\pi_1=\beta_0$ have the expected signs and are significantly different from zero. However, neither $\pi_4$ nor $\pi_5$ are significantly different from zero, so that job turnover and search efficiency were not significantly different during the period of wage-price controls. These results were not affected in any important way by the substitution of an alternative variable used by Goodwin and Carlson to represent the controls period.[12] Moreover, this evidence is consistent with the results reported by Daniel Mitchell and Ross Azevedo (1976) that wage-price controls did not produce significant labor market distortions.

The evidence presented thus far is consistent with the proposition that neither wage misperceptions nor the existence of wage-price controls affected labor turnover or job search efficiency. Nevertheless, drift in the turnover or search parameters could have occurred and generated steady-state increases or decreases in the equilibrium unemployment rate. To accommodate this possibility, (12) was modified as follows:

$$(16)\quad \gamma=\gamma_0+\gamma_1(w-w^e)_{-1}+\gamma_3 T,$$

where $T$ is a time trend variable and $\gamma_3 \gtreqless 0$.

---

[9] In making these comparisons, it should be noted that the estimate of the equilibrium unemployment rate in manufacturing refers to the *experienced* labor force and therefore excludes new entrants. During the sample period, new entrant unemployment was approximately 0.7 percent of the economy-wide civilian labor force (see BLS, 1980, Table 39). Thus, an estimate which is conceptually comparable to those presented by Cagan and Barro would be approximately 5.5 percent rather than 4.8 percent.

[10] I am indebted to an anonymous referee for suggesting this possibility.

[11] This result should not be surprising since the presence of a first-order autoregressive term in the expectations generating equation forces the predictions to adjust rapidly to the behavior of observed wages.

[12] Specifically, this alternative variable measures the percentage of all workers receiving new wage contracts during the preceding twelve months which stipulated increases very near or above the 5.5 percent standard while controls were in effect and thus were presumably constrained by controls. Therefore, this variable captures the pervasiveness of wage controls but not the extent to which wages were held down.

Substitution of (16) into (10) yields

$$(17)\quad \Delta E/E_{-1}=\pi_0+\pi_1 X_{-1}b + \pi_2(w-w^e)_{-1}+\pi_3(w-w^e)X_{-1}+\pi_6 T,$$

where $\pi_6=-\gamma_3$. Column III, Table 2, reports the outcome of estimating (17).[13] As before, the estimated coefficients associated with the basic theoretical parameters have the anticipated signs and are significantly different from zero. Once again, the coefficients on the wage misperception variables are not significantly different from zero. However, the estimate of the trend term coefficient ($\pi_6=-\gamma_3$) is positive and significantly different from zero, implying a steady decline in the turnover rate and, correspondingly, a decrease in the equilibrium unemployment rate throughout the sample period. Although it is hazardous to speculate about all of the forces which are being captured by this result, one possibility is the monotonic decline in the relative minimum wage rate ($\$1.60/W$), where \$1.60 is the fixed nominal minimum wage during this period, and $W$ is nominal average hourly earnings (excluding overtime) in manufacturing. Credible evidence on this hypothesis would, however, require consideration of a larger sample period in which the relative minimum wage exhibited greater variability.

## III. Concluding Remarks

I have presented evidence concerning the ability of Keynesian and equilibrium theories to explain cyclical fluctuations in employment. In the context of a simple but well-known model of labor market dynamics, the Keynesian framework was interpreted as assigning to exogenous variations in the volume of labor market contacts a significant role in determining the time path of employment. On the other hand, the equilibrium paradigm was construed as placing importance on the effects of deviations of actual from expected real wages on labor turnover, the efficiency of job search, and, hence, employment and unemployment. Of course, these two hypotheses are not mutually exclusive and could provide complementary accounts of employment changes over the business cycle.

Based upon data from the U.S. manufacturing sector, the empirical results suggested that changes in the number of vacancy contacts rather than wage misperceptions significantly affected employment growth and equilibrium unemployment. This evidence is consistent with the findings reported by Parsons (1973) and Barron concerning the determinants of manufacturing quits and unemployment duration, respectively, but contradicts the more recent results on unemployment duration in manufacturing obtained by Björklund and Holmlund. The evidence presented in this paper provides some additional support, then, for the view that changes in employment are generated primarily by exogenous variations in labor market quantities rather than by the optimal responses of market participants to misperceived pecuniary opportunities.

Two caveats to this interpretation of the empirical results should be mentioned. First, as noted above, the misperceptions hypothesis implies an unambiguously positive effect of real wage surprises on job search efficiency only if the labor supply response to unanticipated real wage offers dominates the effect on labor demand. In the absence of firm evidence on the relative magnitudes of these elasticities, this qualitative prediction depends on an untested, maintained hypothesis. Second, suppose that changes in the number of job vacancies simply reflect the profit-maximizing responses of firms to (known) variations in the wage misperceptions of job seekers. In this situation, estimates of reduced form relationships investigated here (and by Parsons (1973), Barron, and Björklund-Holmlund) may not be able to discriminate sharply between disequilibrium and search-theoretic explanations for cyclical fluctuations in employment.

---

[13] There was no evidence of a drift in the search efficiency parameter nor did the presence of the (statistically insignificant) wage-price control variables affect these qualitative conclusions. Moreover, the unemployment insurance (*UI*) replacement ratio (the ratio of average weekly *UI* benefits to average weekly earnings) did not significantly decrease the job search efficiency parameter $\beta$, as has been widely hypothesized. The details of these additional results are available from the author upon request.

While these tests of the wage misperception hypothesis are informative, it is important to recall that they are based upon a particular construction of the variable which represents the difference between actual and expected real wages. As Joseph Altonji and Orley Ashenfelter (1980) and Seater have remarked, aggregate real wages are very easy to track with ARMA techniques so that there is little scope for substantial misperceptions of wages over time when expectations are modelled in this manner. However, the finding that the residuals from a rather simple model of the aggregate real wage process are essentially white noise casts doubt on the empirical significance of theories which assign to forecasting mistakes the burden of explaining cyclical fluctuations in employment.

In addition to providing tests of the two hypotheses, the empirical results were used to estimate the equilibrium or natural rate of unemployment and the effects of wage-price controls. The equilibrium unemployment rate in manufacturing was computed to be approximately 4.8 percent of the labor force, in contrast to the observed mean unemployment rate of 5.7 percent over the sample period. Tests of hypotheses and quantitative estimates of the natural rate of unemployment were unaffected by the incorporation of variables which proxied the presence of wage-price controls. There was evidence of a downward drift in the job turnover rate which may be attributable in part to the monotonic decline in the relative minimum wage during the sample period.

Of course, the results reported here are specific to one sector (manufacturing) of the U.S. economy, were generated with data from one business cycle, and were analyzed within the context of a particular theoretical model. The extent to which these findings can be generalized to the U.S. economy as a whole or to other countries, time periods, and models remains a question of considerable importance for future research.

## REFERENCES

**Altonji, Joseph and Ashenfelter, Orley,** "Wage Movements and the Labor Market Equilibrium Hypothesis," *Economica*, August 1980, *47*, 217–45.

**Barro, Robert,** "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–26.

______, "A Capital Market in an Equilibrium Business Cycle Model," *Econometrica*, September 1980, *48*, 1393–417.

**Barron, John,** "Search in the Labor Market and the Duration of Unemployment: Some Empirical Evidence," *American Economic Review*, December 1975, *65*, 934–42.

**Björklund, Anders and Holmlund, Bertil,** "The Duration of Unemployment and Unexpected Inflation: An Empirical Analysis," *American Economic Review*, March 1981, *71*, 121–31.

**Cagan, Phillip,** "The Reduction of Inflation and the Magnitude of Unemployment," in W. Fellner, ed., *Contemporary Economic Problems*, Washington: American Enterprise Institute, 1977, 15–52.

**Friedman, Milton,** "The Role of Monetary Policy," *American Economic Review*, March 1968, *58*, 1–17.

**Goodwin, William and Carlson, John,** "Job Advertising and Wage Control Spillovers," *Journal of Human Resources*, Winter 1981, *16*, 80–93.

**Holt, Charles and David, Martin,** "The Concept of Job Vacancies in a Dynamic Theory of the Labor Market," in *The Measurement and Interpretation of Job Vacancies*, New York: Columbia University Press, 1966, 73–110.

**Lipsey, Richard,** "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis," *Economica*, February 1960, *27*, 1–31.

**Lucas, Robert,** "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

______ **and Rapping, Leonard,** "Real Wages, Employment, and Inflation," in E. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970, 257–305.

**MacRae, C. Duncan, and Schweitzer, Stuart,** "Help-Wanted Advertising, Aggregate Unemployment, and Structural Change," *Proceedings* of the 23rd Annual Meeting of the Industrial Relations Research Association, Madison: Industrial Relations Research Association, 1970, 87–96.

**Mitchell, Daniel and Azevedo, Ross,** *Wage-Price Controls and Labor Market Distortions*, Los Angeles: Institute of Industrial Relations, University of California-Los Angeles, 1976.

**Mortensen, Dale,** "A Theory of Wage and Employment Dynamics," in E. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970, 167–211.

______, "Job Matching Under Imperfect Information," in O. Ashenfelter and J. Blum, eds., *Evaluating the Labor Market Effects of Social Programs*, Princeton: Princeton University Press, 1976, 194–238.

**Parsons, Donald,** "Quit Rates Over Time: A Search and Information Approach," *American Economic Review*, June 1973, *63*, 390–401.

______, "Models of Labor Market Turnover: A Theoretical and Empirical Survey," in R. Ehrenberg, ed., *Research in Labor Economics*, Vol. 1, Greenwich: JAI Press, 1977, 185–223.

**Phelps, Edmund,** "Money Wage Dynamics and Labor Market Equilibrium," in his *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970, 124–66.

**Santomero, Anthony and Seater, John,** "The Inflation-Unemployment Trade-Off: A Critique of the Literature," *Journal of Economic Literature*, June 1978, *16*, 499–544.

**Seater, John,** "Testing Search Theories of Aggregate Labor Supply Cycles," mimeo., North Carolina State University, January 1982.

**U.S. Department of Labor, Bureau of Labor Statistics (BLS),** *Employment and Earnings*, May 1974, *20*, Washington.

______, *Handbook of Labor Statistics*, December 1980, Bulletin 2070, Washington.

______, *Employment and Earnings*, various issues.

______, unpublished tabulations.

# Tax Policy, the Interest Elasticity of Saving, and Capital Accumulation: Numerical Analysis of Theoretical Models

*By* OWEN J. EVANS*

Increasing disillusionment with traditional empirical methods is having a major impact on research methodology, which is increasingly turning to simulation analysis based on theoretical models and assumed parameter values. In research on saving and wealth accumulation, there have been several recent studies[1] consisting of simulation of theoretical models based on the life cycle model. A particular focus has been tax policy and capital formation, which is the area of concentration of this paper.

The aim is to analyze in a theoretical simulation context, the impact on saving and wealth of an increase in the after-tax rate of return to saving, generated by a cut in the tax rate on income from capital. This paper suggests that such simulation analysis, while instructive as to the behavior of the class of models examined, and as to the critical channels of influence, does not yield a theoretical presumption in favor of cutting taxes on income from capital as a means of stimulating capital formation.

The development of this literature derives in part from disillusionment with traditional empirical methods; many such papers begin by listing a litany of the problems involved in the usual type of empirical analysis before proceeding to simulation of theoretical structures. The implicit conclusion at times appears to be that because of the daunting problems of standard empirical methodology, the results derived from it can be ignored, and that empirical magnitudes can perhaps be discovered by deductive means. The temptation to think of the precise numerical results derived from theoretical simulation as empirical can be rather strong; it needs to be reemphasized that empirical magnitudes can only be determined by empirical studies.

In Section I, the theoretical structure of a zero bequest life cycle model[2] is outlined, and several numerical simulations of this model are presented. These simulations highlight interesting and nontrivial interactions of theoretical parameters, and also demonstrate the difficulty involved in attempts to develop robust aggregate implications from such a framework. For instance, it is argued that Summers' claim to have developed "a prima facie theoretical case for a high interest elasticity of savings" (1981a, p. 537) does not stand up to sensitivity analysis—a small change in the assumed rate of pure time preference, or in other parametric assumptions, generate otherwise plausible cases in which "the" interest elasticity of aggregate saving is low. The experiments conducted in Section I are partial equilibrium in the sense that the interest rate is changed exogenously, which changes savings and hence the capital stock, but the feedback effect of the capital stock on the interest rate is not considered. These experiments are also uncompensated, since when the interest rate is exogenously changed, there is no offsetting change elsewhere. Alternatively, these partial equilibrium experiments could be regarded as dealing with a cut in the after-tax interest rate, generated by reducing capital income taxes, and financed by equivalent reduction in government consumption expenditure.

---

*International Monetary Fund. This paper derives from chs. 2 and 3 of my doctoral dissertation, completed during a Research Fellowship at the Brookings Institution. Financial support was received from Brookings and the Social Science Research Council. I am grateful to Henry Aaron, Barry Bosworth, Maxim Engers, Irwin Friend, Laurence Kotlikoff, Franco Modigliani, Joseph Pechman, Duncan Ripley, Robert Shiller, Patricia Stack, and my supervisor Albert Ando for comments and guidance on earlier drafts. None of the above institutions or individuals is responsible for the views expressed or any remaining errors.

[1] For example, see James Tobin and Walter Dolde; Betsy White; Lawrence Summers (1978, 1981a,b), G. Moore; Alan Auerbach and Laurence Kotlikoff; and Auerbach, Kotlikoff, and J. Skinner.

[2] It is popular to establish a sharp dichotomy between "life cycle models" and "bequest models." This taxonomy ignores the frequent references to the potential significance of bequests in the writings of the originators of the life cycle hypothesis. A distinction between "zero bequest life cycle models" and "life cycle models incorporating bequests" is more accurate in terms of doctrinal history. See for example, Albert Ando and Franco Modigliani, and Modigliani and Richard Brumberg.

In Section II, a simulation of the transition path from one steady state to another is conducted; it is found that the saving rate must overshoot its new long-run equilibrium value during the transition in order for the wealth-income ratio to reach its new long-run equilibrium. In addition, in a structure with a large positive partial equilibrium interest elasticity of saving, the necessary transitional fluctuations in the saving rate are so large as perhaps scarcely to be plausible.

In Section III, intergenerational transfers are introduced, and further simulations conducted. If such transfers are a significant component of wealth accumulation,[3] then their response to interest rate changes is a key determinant of the interest elasticity of saving and hence is critical in discussion of tax policy and capital formation.

In Section IV, the consequences of reduction or elimination of taxation of capital income taxes (with compensating increases in some other tax) are examined in a general equilibrium context. Again results are critically dependent on parameter choice.

In general, this type of simulation exercise is seen to be sensitive to parametric structure, and to what factors are included or excluded. Such analysis can be instructive as the key theoretical influences, but *cannot* substitute for empirical research.

## I

The utility function is assumed to take some simple functional form, namely, to belong to the constant proportional risk aversion class:[4]

$$(1)\qquad U=\sum_{t=1}^{T} U(C_t)/(1+\rho)^{t-1}$$

where $U(C_t)=(1/(1-\gamma))C_t^{1-\gamma}$, for $\gamma \neq 1$ or $U(C_t)=log(C_t)$, for $\gamma=1$ and where $T=$ length of (economic) life, known with certainty; $\rho=$ rate of pure time preference; and $\gamma=$ coefficient of proportional risk aversion.[5]

It is further assumed that the individual receives no inheritance and leaves no bequest, that the real interest rate is constant and certain, and that individual labor income grows at some constant rate $g$ per period until retirement, after $T_1$ years in the work force. This yields a budget constraint of the following form:

$$(2)\qquad \sum_{t=1}^{T} C_t/(1+r)^{t-1} = w_1 \sum_{t=1}^{T_1} [(1+g)/(1+r)]^{t-1},$$

where $w_1$ is the individual's wage in the first period of economic life.

The first-order conditions yield the following result:

$$(3)\qquad C_{t+1}=[(1+r)/(1+\rho)]^{1/\gamma}C_t.$$

This defines the slope of the individual consumption profile; (2) and (3) can be solved jointly for the value of $C_1$, after which everything else follows.

In order to make aggregate computations, further assumptions are needed (specifically, that there are $T$ overlapping cohorts alive at any time point, differing only in age, that all individuals in the work force at a given time are paid the same wage, irrespective of age differences, and that population grows at a steady rate of $n$ percent per annum). With this additional structure, one can aggregate over cohorts and examine aggregate magnitudes. The assumptions of the model structure are so highly stylized that it should scarcely need mentioning that attempts to draw policy conclusions may not be robust because of the absence from the structure of such phenomena as uncertainty (of any kind), heterogeneity, and money, to mention only three.

---

[3] Kotlikoff and Summers (1981) suggest that 70 to 80 percent of household wealth in the United States derives from intergenerational transfers.

[4] All the studies mentioned above use this function, which also exhibits a constant elasticity of substitution between consumption in different periods, equal to $1/\gamma$. This is the only *additive* utility function consistent with the Modigliani-Brumberg homogeneity hypothesis (see Menahem Yaari). Robert Merton has shown that the portfolio allocation problem and the consumption/saving decision are separable when the utility function is of this form.

[5] To avoid confusion, it should be noted that what Summers calls $\gamma$ in his paper is equal to $1-\gamma$ in my notation.

Plausible ranges for parameters and results need to be discussed at the outset. For the United States, the annual rates of population and productivity growth (measured in terms of real *GNP* per man-hour) have been 1.4 percent and 2.2 percent over 1900 to 1974. A measure of the portfolio-weighted real after-tax rate of return averaged 4.5 percent over 1900–74.[6] Following Summers (1981a), for most of the simulations I assume a productivity growth rate of 2 percent per annum, population growth rate of 1.5 percent per annum, after-tax real rate of return of 4 percent per annum, an economic lifetime of fifty years, and an earning span of forty years.

The intertemporal elasticity of substitution in consumption $(1/\gamma)$ is more problematic. Warren Weber (1970, 1975) found ranges of 0.13 to 0.41 in his first paper, and 0.56 to 0.75 in his second. Irwin Friend and Marshall Blume (1975) found evidence suggestive of an elasticity of less than 0.5. Sanford Grossman and Robert Shiller (1981) suggested less than 0.25; Robert Hall (1981) found values less than 0.1, and Summers (1981b) found values from 0.04 to 0.45. The present simulations use values of $\gamma$ ranging from 1.0 to 5.0 (i.e., the intertemporal elasticity ranges from 0.2 to 1.0).

Because of the absence of evidence as to time preference rates, the assumptions are varied, ranging from $-3$ percent to 3 percent.[7] Negative time preference rates cannot be ruled out a priori. Plausible ranges of results for saving rates and wealth income ratios in the United States can be established. Household wealth-income ratios[8] averaged around 4.0 over the 1970's, and household saving rates measure around 10 percent, depending on the precise definition employed.

In Table 1, the results of some simulations are provided, over a variety of values for $\gamma$ and for the rate of time preference. Statistics presented are the interest elasticity of saving, the saving rate, and the wealth-income ratio. The interest elasticities of saving computed are partial equilibrium, as the feedback effect of the capital stock on the interest rate is not considered. If the increase in the rate of interest is viewed as deriving from a cut in the tax on capital income, then the elasticities computed are uncompensated, as no offsetting change is made elsewhere.

It can be seen in Table 1 that the higher is the value of the risk aversion coefficient $\gamma$, the lower is the interest elasticity, the saving rate, and the wealth-income ratio. The interest elasticity is lower because $1/\gamma$ is the intertemporal elasticity of substitution of consumption. The saving rate and wealth-income ratios both decline with $\gamma$, because the larger is $\gamma$, the flatter is the individual consumption profile (see equation (3)). A relatively flat consumption profile and rising labor income may generate substantial *dissaving* in youth lowering the saving rate and wealth-income ratio. It can also be seen that the larger is the time preference rate $\rho$, the lower are the saving rate and wealth-income ratio and the higher is the interest elasticity of saving. As can be seen from equation (3), the higher is the pure time preference rate, the flatter is the lifetime consumption profile of the individual, so that saving in youth is smaller, or dissaving in youth greater. It is not clear that substantial dissaving in youth (i.e., generation of significantly negative net worth) is empirically realistic. With dissaving

---

[6] See Kotlikoff and Summers (1981, pp. 714–15), for the origin of these estimates.

[7] However, for some inferential evidence, see Mancur Olson and Martin Bailey (1981).

[8] Using data from the MPS model data bank for household wealth and disposable income.

TABLE 1—INTEREST ELASTICITIES OF SAVING, SAVING RATES, AND WEALTH-INCOME RATIOS

| | (1): $\rho = -0.03$ | | | (2): $\rho = 0.00$ | | | (3): $\rho = +0.03$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Value of $\gamma$ | Saving Elasticity | Saving Rate | Wealth-Income | Saving Elasticity | Saving Rate | Wealth-Income | Saving Elasticity | Saving Rate | Wealth-Income |
| 1.0 | 0.70 | 0.31 | 9.17 | 1.11 | 0.34 | 9.83 | 3.55 | 0.06 | 1.64 |
| 2.0 | 0.48 | 0.19 | 5.46 | 0.74 | 0.11 | 3.25 | 3.24 | 0.04 | 1.09 |
| 3.0 | 0.41 | 0.15 | 4.49 | 0.56 | 0.08 | 2.21 | 2.97 | 0.03 | 0.82 |
| 4.0 | 0.34 | 0.13 | 3.81 | 0.40 | 0.06 | 1.67 | 1.90 | 0.01 | 0.40 |

*Notes:* Assumptions: $T = 50$; $T_1 = 40$; $g = .02$; $n = .015$; $r = .04$. The saving elasticity is the interest elasticity of saving, computed numerically as the proportionate change in aggregate saving, for a given (small) proportionate change in the rate of interest. Saving rates and wealth-income ratios are computed relative to total income (including interest).

TABLE 2—INTEREST ELASTICITIES, SAVING RATES, AND WEALTH-INCOME RATIOS

| | (1): $\rho = -0.03$ | | | (2): $\rho = 0.00$ | | | (3): $\rho = +0.03$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Value of $\gamma$ | Saving Elasticity | Saving Rate | Wealth-Income | Saving Elasticity | Saving Rate | Wealth-Income | Saving Elasticity | Saving Rate | Wealth-Income |
| 1.0 | 0.72 | 0.21 | 10.47 | 0.90 | 0.15 | 7.40 | 1.91 | 0.06 | 3.16 |
| 2.0 | 0.46 | 0.13 | 6.84 | 0.56 | 0.09 | 4.71 | 1.04 | 0.05 | 2.37 |
| 3.0 | 0.30 | 0.10 | 5.24 | 0.35 | 0.07 | 3.70 | 0.61 | 0.04 | 2.10 |
| 4.0 | 0.19 | 0.09 | 4.38 | 0.23 | 0.06 | 3.18 | 0.36 | 0.04 | 1.97 |
| 5.0 | 0.11 | 0.08 | 3.84 | 0.13 | 0.06 | 2.86 | 0.20 | 0.04 | 1.89 |

*Notes:* Assumptions: $T = 50$; $T_1 = 40$; $g = .01$; $n = .01$; $r = .04$.

in youth, the income effect of an interest rate change works in the opposite direction than that usually imagined, as a higher interest rate implies higher interest outlays, not higher interest receipts. This is the reason for the large interest elasticities in column (3) of Table 1 (which corresponds to column (1) of Summers, 1981a, Table 1, p. 536). In the cases examined in columns (1) and (2) of Table 1, there is no dissaving in youth, and the interest elasticities are sharply lower.

The results demonstrate the extreme sensitivity of Summers' "prima facie theoretical case" for a large interest elasticity of savings to changes in the time preference rate. We see one case in column (2) in which the elasticity of saving is 0.7 (still larger than most empirical estimates, but sharply lower than the figure of around 2.0 proposed by Summers), the saving rate is 11 percent and the wealth-income ratio is 3.25—well within plausible empirical ranges. In column (1) of Table 1, we see a case ($\gamma = 4.0$), in which the saving rate is 13 percent, the wealth-income ratio 3.8, and the interest elasticity of saving 0.34, which is dramatically less than the figure of around 2.0 proposed by Summers.

It is well known that productivity growth and population growth in the United States have diminished sharply in the 1970's. In Table 2, simulation results are presented in which the productivity and population growth rates are assumed each to be 1 percent per annum.

At the individual level, the lower is the assumed productivity growth rate, the smaller is the "reduction of human wealth" effect, emphasized by Summers (1981a). When the rate of interest changes, the substitution and reduction of human wealth effects dominate in the younger cohorts, causing increased saving, and the income effect dominates in the older cohorts, reducing saving.[9] Computation of the aggregate interest elasticity of saving is therefore extremely sensitive to the population and income growth assumptions. The smaller the assumed rates of population and productivity growth, the greater the relative size of the older cohorts (numerically, and in terms of wealth), the larger is the aggregate income effect, and the smaller is the interest elasticity of saving.

---

[9] For more detail, see my dissertation, ch. 3.

The results in Table 2 clearly illustrate these considerations. If we restrict ourselves to nonnegative rates of time preference, we see in column (2), a case ($\gamma = 3.0$) in which the saving rate and wealth-income ratio are within plausible ranges, and the interest elasticity is 0.35—far below the figure of around 2.0 advocated by Summers. If negative time preference rates are regarded as admissible (and they cannot be dismissed a priori), we see an otherwise plausible case ($\gamma = 5.0$) with an interest elasticity of 0.1 in column (1). Clearly the "prima facie theoretical case for a high interest elasticity of savings" proposed by Summers is not robust to changes in assumptions.

Two factors not present in the above structure are intergenerational transfers and uncertainty as to the future labor income stream. In Section III below, it will be argued that the presence of bequests is likely further to reduce the aggregate interest elasticity of saving, compared to the zero bequest case. Labor income uncertainty is not incorporated into the models in this paper, but the direction of the effect would likely be to reduce the simulated interest elasticity of saving. It has been shown in several contexts[10] that a sufficient condition for an increase in labor income uncertainty to increase saving is that the utility function exhibit diminishing absolute risk aversion—a relatively weak property possessed by the constant proportional risk aversion utility function used here. With increased saving, a given change in saving would translate into a smaller elasticity. Further, the saving rates and wealth-income ratios would be larger, and closer to plausible ranges, for larger values of $\gamma$—for which the simulated interest elasticity would be small. Hence incorporation of labor income uncertainty would be likely to strengthen the plausibility of a "small" interest elasticity of saving.

## II

The above discussion of the interest elasticity of aggregate saving was entirely in terms of comparison of steady states, without any analysis of the transition. For policy purposes, the transition may be of as much importance as the steady-state analysis.

The elements of the problems are developed as follows. The change in the interest rate is a surprise. Before its occurrence, all lifetime calculations have been made using the initial interest rate. After a change in the rate, it is assumed that the new rate will persist forever, and all cohorts instantaneously recompute their lifetime wealth (recalculating the present value of future labor income and adding it to wealth accumulated) and calculate a new optimal consumption profile. There are no expectational lags, long-term contracts, or habit persistence in consumption.

Nevertheless, the adjustment of the aggregate system from its old steady-state path to the new steady state does not occur in one period, but rather is gradual and only complete when all cohorts alive at the time of the change have died.[11] This is because the adjustment of the aggregate system has two components—the first is the adjustment and recalculation made by the transitional cohorts in the period following the change; the second is the gradual dying out of these transitional cohorts, and their replacement by new generations for whom the *ex post* regime was present at the beginning of their economic lives.

In Table 3, numerical results are presented for the transition from steady state *ex ante* to steady state *ex post* for a zero bequest model, as outlined in Section I, after a rise in the

---

[10] See Jacques Drèze and Modigliani (1972), Hayne Leland (1968), Bruce Miller (1974, 1976), and Agnar Sandmo (1970).

[11] In a model with bequests, the convergence from the steady state *ex ante* to the steady state *ex post* might be an even more gradual process, as the effects of the initial disturbance would be transmitted to newly born generations through the inheritances received from cohorts alive at the time of the change.

TABLE 3—TRANSITION FROM ONE STEADY STATE TO ANOTHER AFTER A CHANGE IN THE INTEREST RATE

| | Transition after a rise in the interest rate from 4 percent per annum to 6 percent per annum | | | Transition after a fall in the interest rate from 6 percent per annum to 4 percent per annum | | |
|---|---|---|---|---|---|---|
| Year | Saving Rate | $C$-$YL$ Ratio | Wealth-Income Ratio | Saving Rate | $C$-$YL$ Ratio | Wealth-Income Ratio |
| 0 | 0.056 | 1.008 | 1.645 | 0.133 | 1.121 | 3.920 |
| 1–5 | 0.208 | 0.894 | 2.114 | −0.042 | 1.223 | 3.658 |
| 6–10 | 0.198 | 0.940 | 2.643 | −0.021 | 1.164 | 3.032 |
| 11–15 | 0.188 | 0.980 | 3.045 | −0.004 | 1.119 | 2.557 |
| 16–20 | 0.179 | 1.013 | 3.349 | +0.011 | 1.084 | 2.210 |
| 21–25 | 0.170 | 1.043 | 3.573 | 0.023 | 1.059 | 1.968 |
| 26–30 | 0.161 | 1.068 | 3.730 | 0.035 | 1.039 | 1.810 |
| 31–35 | 0.152 | 1.089 | 3.832 | 0.043 | 1.025 | 1.716 |
| 36–40 | 0.144 | 1.105 | 3.890 | 0.050 | 1.016 | 1.667 |
| 41–45 | 0.137 | 1.115 | 3.914 | 0.054 | 1.010 | 1.649 |
| 46–50 | 0.134 | 1.120 | 3.920 | 0.056 | 1.008 | 1.645 |
| 51 | 0.133 | 1.121 | 3.920 | 0.056 | 1.008 | 1.645 |

*Notes:* Assumptions: $T = 50$; $T_1 = 40$; $g = 0.02$; $n = 0.015$; $\gamma = 1.0$; $\rho = 0.03$. The saving rate and wealth-income ratio are measured in terms of total income (including interest income); $C$-$YL$ is the consumption-labor income ratio; period 0 is the steady state *ex ante* and period 51 is the steady state *ex post*.

interest rate from 4 to 6 percent and for completeness, after a fall in the interest rate from 6 to 4 percent. For ease of presentation, the figures are presented as averages over five-year periods. Period 0 is the steady state *ex ante*, period 51 is the steady state *ex post*.

What we observe is a dramatic *overshooting* by the saving rate immediately following the exogenous change. In the case of an interest rate rise, the saving rate jumps from 5.6 to 20.8 percent as the transitional cohorts make their optimal adjustments. Subsequently, the saving rate declines from the value to which it overshot toward its ultimate new steady state position. This type of overshooting behavior by the saving rate is necessary, because the wealth-income ratio has to increase substantially during the transition. In the case of a decline in the interest rate, the saving rate overshoots dramatically resulting in a *negative* saving rate for the first *fifteen years* after the change. Such overshooting is again necessary because during this transition, a large fraction of accumulated wealth has to be consumed.

Such huge movements in the aggregate saving rate during the transition are necessary for the *long-run* interest elasticity of saving to be in the range suggested by Summers (1981a), and cast further doubt on his argument, as movements of aggregate saving rates anywhere near this magnitude do not appear to have been observed historically.

## III

In this section, bequests are incorporated into the structure. One way to incorporate bequests is for the utility or consumption of children to be an argument of the utility function of parents (see Robert Barro, 1974). In this context, the representative individual acts as if he will live forever[12] (since the utility or consumption of all his descendants enters his own utility function), and as noted by Summers (1981a), the partial equilibrium interest elasticity of saving is infinite, as the after-tax rate of interest is always driven to a modified Golden Rule position.

---

[12] For a discussion of the plausibility of this structure, see Willem Buiter (1979).

In this context, the long-run general equilibrium capital income tax rate[13] elasticity of saving is entirely determined by the production technology. Some illustrative calculations are provided below. The production technology is assumed to be Cobb-Douglas. That is,

$$y = k^{\alpha}, \tag{4}$$

where $y$ is output per (effective) worker, and $k$ is capital per (effective) worker, and $\alpha$ is the capital share of output. The gross return to capital is given by

$$R = \alpha k^{\alpha-1}, \tag{5}$$

and the net return to capital by

$$r_n = (1-\tau_k)\alpha k^{\alpha-1}, \tag{6}$$

where $\tau_k$ is the (assumed proportional) tax rate on capital income. Since the after-tax interest rate is driven to a modified Golden Rule position and labor supply is given, the long-run elasticity of the capital stock (and hence saving) can be found by assuming $r_n$ constant and manipulating equation (6). The resulting expression is

$$\frac{d\,ln(k)}{d\,ln(\tau_k)} = \frac{-\tau_k}{(1-\tau_k)(\alpha-1)}. \tag{7}$$

Using the parametric assumptions employed by Summers (1981a) (i.e., a capital share of output of 0.25, and a tax rate on capital income of 0.5), (7) yields an elasticity of saving of 1.333.[14] That is, an *infinite* partial equilibrium elasticity of saving reduces to a general equilibrium elasticity of 1.333, given the production assumptions made. Presumably, therefore, given these assumptions (made by Summers, 1981a), 1.333 is an *upper bound* on the general equilibrium elasticity, so that a figure of close to 2.0, is out of the question.

There are other, perhaps more plausible, ways of incorporating bequests, than assuming infinite horizon Barro-type dynasties. Two alternatives are discussed briefly below. The individual receives an inheritance, and must pass on at his death a bequest related by the appropriate steady-state growth rate to his initial inheritance.[15] In the first alternative examined, bequests are analogous to final period consumption, and so are maintained proportionate to it. In the second alternative, inheritances and bequests are invariant to changes in the interest rate; this is a convenient benchmark case. In both instances, the inheritance is received at the start of economic life, and the bequest passed on at the end of life. The rest of the structure is as before.

Results are presented in Table 4. In general, the larger the inheritance and bequest, the larger the saving rate and wealth-income ratio, and the smaller the interest elasticity of saving. The interest elasticity is lower for two main reasons—first, with a larger saving rate, a given change in saving translates into a smaller elasticity, and second (in the case in which the inheritance and bequest are invariant to the interest rate), the net income from the bequest increases, partially offsetting the "reduction in human wealth" effect. The addition of intergenerational transfers to the structure also makes small or negative interest elasticities of saving more plausible, by increasing the saving rate and wealth-income ratio for high values of $\gamma$ for which the interest elasticity is bound to be low. The interest elasticities in Table 4 once again reinforce the plausibility of a small interest elasticity of saving, and introduce the possibility of a *negative* elasticity.

The broad conclusions are that incorporation of intergenerational transfers makes a small interest elasticity of saving even more plausible, and enables negative interest elasticities of saving to become entirely plausible. In the convenient benchmark case in which intergenerational transfers are unaffected by the interest rate, several otherwise

---

[13] Also in this context, the interest elasticity and the tax rate elasticity of saving differ, with the latter being the interesting concept for the purposes at hand.

[14] A value of $\tau_k$ of 0.25 would imply an elasticity of 0.444; a value of $\tau_k$ of 0.1 would imply an elasticity of 0.15.

[15] See my dissertation, ch. 3, for details.

TABLE 4—INTEREST ELASTICITIES, SAVING RATES, AND WEALTH-INCOME RATIOS IN MODELS WITH INTERGENERATIONAL TRANSFERS

| Value of $\gamma$ | Inheritance Ratio | Interest Elasticities of Savings (1) | Interest Elasticities of Savings (2) | Saving Rate (3) | Wealth-Income Ratio (4) |
|---|---|---|---|---|---|
| 1.0 | 0.0 | 1.25 | 1.25 | 0.16 | 4.69 |
| | 1.0 | 1.25 | 0.96 | 0.20 | 5.73 |
| | 2.0 | 1.28 | 0.78 | 0.23 | 6.67 |
| | 3.0 | 1.31 | 0.65 | 0.26 | 7.53 |
| 3.0 | 0.0 | 0.75 | 0.75 | 0.06 | 1.66 |
| | 1.0 | 0.53 | 0.29 | 0.10 | 2.99 |
| | 2.0 | 0.50 | 0.13 | 0.14 | 4.18 |
| | 3.0 | 0.51 | 0.04 | 0.18 | 5.25 |
| 5.0 | 0.0 | 0.37 | 0.37 | 0.03 | 1.00 |
| | 1.0 | 0.20 | −0.01 | 0.08 | 2.39 |
| | 2.0 | 0.20 | −0.10 | 0.12 | 3.64 |
| | 3.0 | 0.24 | −0.14 | 0.16 | 4.76 |

*Notes:* Assumptions: $T = 50$; $T_1 = 40$; $\rho = 0.01$; $g = 0.02$; $n = 0.015$; $r = 0.04$. Column (1) keeps the bequest proportional to final period consumption and Column (2) has inheritances and bequests invariant to the interest rate. An inheritance ratio of 1.0 means an inheritance equal to the wage in the first year of economic life.

plausible cases are developed in which the interest elasticity of saving is negative. The key channel of influence, when inheritances and bequests are important, is the effect on them of interest rate changes.

It is also possible for bequests to enter the analysis *in the absence of a bequest motive.* If one assumes that length of life is stochastic, and that actuarially fair life insurance and annuities markets are absent, the individual bears all the risk from his stochastic lifetime. (See my dissertation, ch. 3, for a detailed analysis.) Because the risk of living longer than expected and running out of wealth outweighs that of dying too soon with some wealth unconsumed, bequests are substantially positive in this framework on average, despite the complete absence of a bequest motive.[16] The induced bequests can be quite large and the interest elasticity of aggregate saving is frequently *negative* in this context. This framework suggests that it may be problematic to attempt to infer a significant *bequest motive* from the (alleged) empirical significance of bequests.[17]

## IV

The above analysis focussed on the partial equilibrium interest elasticity of savings. The rate of interest was purely an exogenous parameter. In this section, the analysis is extended to the general case, in which a (tax-induced) interest rate change affects savings, thus changing the capital stock, which in turn feeds back to the interest rate. This is done by adding a production function to the structure.

Production is assumed to be of the Cobb-Douglas type, and gross returns are equated to marginal products. That is,

---

[16] James Davies conducts a similarly motivated analysis in an attempt to reconcile the findings from survey evidence which indicate that wealth decumulation in retirement is markedly less than predicted by the zero bequest model, but that respondents rarely cite bequests as a motivation for saving.

[17] See Kotlikoff and Summers, and Michael Darby. This apparently arcane question is of some theoretical significance. If we have substantial bequests without (or with a small) bequest motive, then we would not expect a significant offset from intergenerational transfers (see Barro) on the effect of Social Security on personal saving. In many areas, important results depend not only on the empirical significance of bequests, but also on their deriving from a bequest motive.

$$Y_t = K_t^{\alpha}\left((1+g)^t L_t\right)^{1-\alpha}, \tag{8}$$

where $Y$ is gross output, $K_{t-1}$ is the capital stock at the start of the given period, and $L$ is the stock of labor. One derives from the above,

$$R_t = \alpha(Y_t/K_t), \tag{9}$$

$$W_t = (1-\alpha)(Y_t/L_t), \tag{10}$$

where $R$ and $W$ are the gross returns.

The net returns which enter the household calculus are given by

$$r_n = R(1-\tau_k), \tag{11}$$

$$w_n = W(1-\tau_w), \tag{12}$$

where $\tau_k$ and $\tau_w$ are the proportional tax rates on income from capital and labor income, respectively.

Two cases are examined in detail. In each case, the aggregate saving rates and wealth-income ratios are broadly similar, but in the first case, the elasticity of substitution in consumption is high and there are no bequests, and in the second case, the elasticity of substitution in consumption is low and there are substantial bequests.

In the first instance, each individual has a *log* utility function, with other assumptions as in Table 4.[18] These yield a personal saving rate of 16 percent, and a wealth-disposable income ratio of 4.7. In the second instance, each individual has a proportional risk-aversion coefficient of 5.0 (alternatively, an intertemporal elasticity of substitution in consumption of 0.2), receives a bequest in the first year of economic life equal to three times his wage in that year, and passes on the appropriate bequest fifty years later. Inheritances and bequests remain constant under changes in the interest rate. These assumptions yield a personal saving rate of 16 percent and a wealth-income ratio of 4.8.

The experiments conducted are in the differential incidence tradition of public finance. The after-tax rate of interest is increased by a cut in the tax on capital income. Government revenue is maintained at the initial level[19] by raising either the wage tax or the consumption tax.[20]

The first experiment was to induce a small change in the interest rate in order to compare the general equilibrium with the partial equilibrium response. In the first case,[21] the general equilibrium response should be substantially *smaller* than the partial equilibrium response, as the increase in savings will increase the capital stock, and thus reduce the gross and net returns to capital. The degree of damping of the partial equilibrium effect is substantial—the increase in saving generated by a small cut in capital income taxes, financed by a rise in the consumption tax rate, is 68 percent of the partial equilibrium increase. When the compensatory tax increase was through the wage tax, the general equilibrium saving increase is only 54 percent of the partial equilibrium increase.

In the second case, with a negative partial equilibrium interest elasticity, the changes in the capital stock amplify rather than damp the partial equilibrium change. The increase

---

[18] That is, economic life of length fifty years, earning span of forty years, time preference rate of 1 percent per annum, initial interest rate of 4 percent, population growth rate of 1.5 percent, and (labor augmenting) technological progress at 2.0 percent.

[19] The sense in which these are compensated changes needs to be discussed. Government revenue is kept at the same level, but the income effects on the representative individual may still be substantial, so that we are not looking at a pure substitution effect by any means.

[20] The results are derived by simulation using a modified Gauss-Seidel iteration technique. Starting from the initial steady state whose structure is known, the tax rate on capital income is reduced, and the wage tax or consumption tax increased by just enough to keep government revenue constant, if nothing else changes. A second simulation is conducted with these tax rates and new after-tax returns (on labor and capital) as inputs, generating new consumption, saving, and wealth results. Gross returns are regenerated from the production function, and the wage or consumption tax is altered again so as to keep government revenue constant. These act as inputs for the next round. The process is iterated to convergence.

[21] The first case ($\gamma = 1.0$, no bequests) has a partial equilibrium interest elasticity of saving of 1.25. The second case ($\gamma = 5.0$, large bequests) has a partial equilibrium interest elasticity of $-0.14$. See Table 4.

TABLE 5—GENERAL EQUILIBRIUM EFFECTS OF ELIMINATION OF THE CAPITAL INCOME TAX

| New Regime | Percentage Change in: Consumption | Wealth | Saving | Final Interest Rate | Wage Tax Rate | Consumption Tax Rate | Initial Saving Rate | Final Saving Rate |
|---|---|---|---|---|---|---|---|---|
| Case 1 | | | | | | | | |
| Consumption Tax | +67.2 | +158.5 | +158.5 | 4.92 | 0.0 | 21.6 | 16.00 | 22.75 |
| Wage Tax | +48.8 | +101.0 | +101.0 | 5.62 | 26.7 | 0.0 | 16.00 | 20.46 |
| Case 2 | | | | | | | | |
| Consumption Tax | +0.2 | +0.9 | +0.9 | 7.95 | 0.0 | 31.6 | 16.2 | 16.3 |
| Wage tax | −3.0 | −12.3 | −12.3 | 8.82 | 39.3 | 0.0 | 16.2 | 14.9 |

*Notes:* Assumptions: $T = 50$, $T_1 = 40$, $\rho = 0.01$, $r = 0.04$, $g = 0.02$, $n = 0.015$. Case 1: $\gamma = 1.0$, no bequests. Case 2: $\gamma = 5.0$; individual in initial steady state receives inheritance equal to three times his initial net wage in the start of economic life.

in the after-tax interest rate reduces saving, and hence the capital stock, and so increases the interest rate further. In the case of a small cut in capital income taxes, financed by a small increase in the consumption tax, the general equilibrium *fall* in saving was 118 percent of the partial equilibrium fall. When the compensating tax increase was through the wage tax, the general equilibrium response was 156 percent of the partial equilibrium response. This suggests that in interpreting earlier analysis of partial equilibrium interest elasticities, one should be aware that when the partial equilibrium interest elasticity of saving is positive, the corresponding general equilibrium elasticity is likely to be substantially smaller, but when the partial equilibrium elasticity is negative, the corresponding general equilibrium elasticity is likely to be substantially greater (in absolute value).

The second set of experiments involved large changes. In one test, the wage tax and capital income tax were eliminated, and replaced by a consumption tax. In the second test, the capital income tax was eliminated, and all government revenue raised from a wage tax. Results are presented in Table 5.

As before, we have two parametric structures—one with a high elasticity of substitution in consumption and no intergenerational transfers, and the other with a low intertemporal elasticity of substitution in consumption, and significant intergenerational transfers.

In the first case, the results resemble closely those of earlier simulation studies. A complete shift to consumption taxation dramatically increases steady-state consumption (by 67 percent), and the capital stock (by 159 percent). In the case of a complete shift to wage taxation, steady-state consumption rises by 49 percent, and the capital stock more than doubles. The reason for the more substantial results under a shift to consumption taxation is that such a system postpones tax payments, thus reducing their present value at the start of the individual's economic life, and so effectively increases his lifetime resources.

In the second case, the results are dramatically different from those of earlier studies. In the case of the shift to a consumption tax, steady-state consumption rises by 0.2 percent, and the steady-state capital stock by 0.9 percent. The final interest rate is 7.95 percent. Essentially, the effect of a consumption tax in reducing the lifetime present value of taxation (by postponing tax payments to later in life) very slightly outweighs the effect of the negative partial equilibrium interest elasticity. Of course, this "tax postponement effect" means that during the transition, some cohorts pay wage and capital income taxes early in life under the old regime, and consumption taxes late in life under the new regime, and so suffer significant reductions in welfare.

When capital income taxes are abolished by a shift to wage taxation, steady-state consumption *falls* by 3 percent, and the steady-state capital stock *falls* by 12 percent. The after-tax interest rate (4 percent in the steady state *ex ante*, and 8 percent immediately after elimination of the 50 percent tax on capital income) is 8.8 percent in the steady state *ex post*, as the preference structure and technology interact in such a way as to amplify the initial effect.

Given these results, it seems unlikely that wage and consumption taxes could be combined to yield a Pareto superior result compared to the steady state, *ex ante*. It would appear from this second case, that the alleged theoretical presumption[22] in favor of elimination of taxes on capital income as an appropriate way to stimulate capital formation, and their replacement by wage or consumption taxes, does not necessarily extend to cases in which bequests generate a significant fraction of capital formation. In the case examined here, in which inheritances and bequests remained constant after the post-tax interest rate changed, the character of the results was dramatically different from the zero bequest case. How inheritances and bequests are likely to change after interest rate changes is something that economists know very little about at present, either theoretically or empirically. If intergenerational transfers play the major role in capital accumulation,[23] then their response to (tax-induced) interest rate changes should be the key point of focus in analysis of tax policy and capital formation.

## V

The life cycle model of household saving was shown to be consistent with values for the interest elasticity of saving ranging from close to zero on one hand, to very large on the other. When intergenerational transfers were introduced, negative interest elasticities of saving became entirely plausible. Attempts made in earlier studies to show that realistically formulated life cycle models necessarily generate large positive interest elasticities of saving were shown to derive from arbitrary choice of parameters, and consequently were not robust to small changes in assumptions.

When the transition from one steady state to another was examined, it was found, as a predictable consequence of stock-flow interactions, that if the long-run saving rate and wealth-income ratio were to rise, the short-run saving rate would have to overshoot its long-run destination, and then decline. It was suggested that if the long-run interest elasticity of saving were as large as proposed in some studies (greater than 2.0), the necessary transitional fluctuations in the saving rate would be huge, perhaps beyond the bounds of historical experience.

When general equilibrium behavior was examined, by addition of a production function, it was found that elimination of the tax on income from capital, and its replacement by a wage or consumption tax, could have large, negligible, or perverse effects on capital formation, depending on the assumed parameters, and the extent to which intergenerational transfers were incorporated. Consequently, unambiguous conclusions could not be drawn.

The major methodological implication is that simulation of theoretical models based on assumed parameters may be better suited for illustrative exercises, than in attempts to provide clear-cut theoretical results, for it is difficult to determine whether a particular result is in fact general, or is a consequence of the choice of parameters. Consequently, claims of generality should be treated with caution unless substantial sensitivity tests have been conducted.

## REFERENCES

**Ando, Albert and Modigliani, Franco,** "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *American Economic Review*, March 1963, *53*, 55–84.

**Auerbach, Alan and Kotlikoff, Laurence,** "National Savings, Economic Welfare, and the

---

[22] This is proposed by Summers (1981a) and Auerbach and Kotlikoff (1981).

[23] As argued by Kotlikoff and Summers (1981).

Structure of Taxation," in M. Feldstein, ed., *Behavioral Simulation Methods in Tax Policy Analysis*, forthcoming.

______, ______, **and Skinner, J.,** "The Efficiency Gains from Dynamic Tax Reform," *International Economic Review*, February 1983, *24*.

**Barro, Robert,** "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

**Buiter, Willem,** "Government Finance in an Overlapping Generations Model with Gifts and Bequests," in G. von Furstenberg, ed., *Social Security versus Private Saving*, Cambridge: Ballinger, 1979.

**Darby, Michael,** *The Effects of Social Security on Income and the Capital Stock*, Washington: American Enterprise Institute, 1979.

**Davies, James,** "Uncertain Lifetime, Consumption, and Dissaving in Retirement," *Journal of Political Economy*, June 1981, *89*, 561–77.

**Drèze, Jacques and Modigliani, Franco,** "Consumption Decisions under Uncertainty," *Journal of Economic Theory*, December 1972, *5*, 308–35.

**Evans, Owen,** "The Life Cycle Inheritance: Theoretical and Empirical Essays on the Life Cycle Hypothesis of Saving," unpublished doctoral dissertation, University of Pennsylvania, 1982.

**Friend, Irwin and Blume, Marshall,** "The Demand for Risky Assets," *American Economic Review*, December 1975, *65*, 900–22.

**Grossman, Sanford and Shiller, Robert,** "The Determinants of the Variability of Stock Prices," *American Economic Review Proceedings*, May 1981, *71*, 222–27.

**Hall, Robert E.,** "Intertemporal Substitution in Consumption," Working Paper No. 720, National Bureau of Economic Research, July 1981.

**Kotlikoff, Laurence and Summers, Lawrence,** "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

**Leland, Hayne,** "Saving and Uncertainty: The Precautionary Demand for Saving," *Quarterly Journal of Economics*, August 1968, *82*, 465–73.

**Merton, Robert,** "Lifetime Portfolio Selection Under Uncertainty: The Continuous Time Case," *Review of Economics and Statistics*, August 1969, *51*, 247–57.

**Miller, Bruce,** "Optimal Consumption with a Stochastic Income Stream," *Econometrica*, March 1974, *42*, 253–66.

______, "The Effect on Optimal Consumption of Increased Uncertainty in Labor Income in the Multiperiod Case," *Journal of Economic Theory*, August 1976, *13*, 154–67.

**Modigliani, Franco and Brumberg, Richard,** "Utility Analysis and Aggregate Consumption Functions: An Attempt at Integration," in A. Abel, ed., *The Collected Papers of Franco Modigliani*: Vol. 2, *The Life Cycle Hypothesis of Saving*, Cambridge: MIT Press, 1980.

**Moore, G.,** "Taxes, Inflation, and Capital Formation," in J. Enzler et al., eds., *Public Policy and Capital Formation*, Washington: Board of Governors of the Federal Reserve System, 1981.

**Olson, Mancur and Bailey, Martin, J.,** "Positive Time Preference," *Journal of Political Economy*, February 1981, *89*, 1–25.

**Sandmo, Agnar,** "The Effect of Uncertainty on Saving Decisions," *Review of Economic Studies*, July 1970, *37*, 353–60.

**Summers, Lawrence,** "Tax Policy in a Life Cycle Model," Working Paper No. 302, National Bureau of Economic Research, November 1978.

______, (1981a) "Taxation and Capital Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, *71*, 533–44.

______, (1981b) "Tax Policy, the Rate of Return and Savings, unpublished paper, December 1981.

**Tobin, James,** "Life Cycle Saving and Balanced Growth," in W. Fellner, ed., *Ten Economic Studies in the Tradition of Irving Fisher*, New York: J. Wiley and Sons, 1967.

______ **and Dolde, Walter,** "Wealth, Liquidity, and Consumption," *in Consumer Spending and Monetary Policy*: *The Linkages*, Federal Reserve Bank of Boston, 1971.

**Weber, Warren,** "The Effect of Interest Rates on Aggregate Consumption," *American*

*Economic Review*, September 1970, *60*, 591–600.

———, "Interest Rates, Inflation, and Consumer Expenditures," *American Economic Review*, December 1975, *65*, 843–58.

**White, Betsy B.**, "Empirical Tests of the Life Cycle Hypothesis," *American Economic Review*, September 1978, *68*, 547–60.

**Yaari, Menahem**, "On the Consumer's Lifetime Allocation Process," *International Economic Review*, September 1964, *5*, 304–17.

# Budget Constraint Prices as Preference Changing Parameters of Generalized Fechner-Thurstone Direct Utility Functions

*By* R. L. BASMANN, D. J. MOLINA, AND D. J. SLOTTJE*

Let $V(X)$ be any direct utility function characterizing consumer preferences that are reflexive, transitive, and complete, and whose parameters are fixed relative to the usual linear budget constraint $p \cdot X = M$. Here $X$ designates the vector $(X_1, \ldots, X_n)$ and is confined to the positive $n$-orthant; $X_i$ designates quantity of the $i$th commodity; $p$ is the vector $(p_1, \ldots, p_n)$ of positive commodity prices; and $M$ designates total expenditure.

Let $X(p, M)$ designate the system of demand functions derived from $V(X)$ subject to $pX = M$, which possess the properties of zero degree homogeneity, symmetry, and semidefinite negativity. Let $W(X; \theta)$ designate a direct utility function with (a) parameter vector $\theta$ dependent on the price vector $p$ that appears in the budget constraint and (b) that rationalizes the same system of demand functions as $V(X)$, namely, $X(p, M)$. Probably, many economists who have considered price-dependent preferences[1] would conjecture that every such price-dependent utility function $W(X; p)$ satisfying (a) and (b) is of the rather restricted form

$$(1) \quad W(X; p) = F(p)V(X) + H(p),$$

or of a form equivalent to (1) under a monotonic transformation (see Robert Pollak, 1977, p. 71). The relation (1) implies that marginal rates of substitution are independent of the price-vector $p$ at every point $X$ of the commodity space, so that the indifference maps of $V(X)$ and $W(X; p)$ coincide everywhere.

In this paper, we provide a simple counterexample that rejects the above-mentioned conjecture. Since only one counterexample is required, we find it convenient to employ explicit closed form direct utility functions.

Let $U(X; \theta)$ be a direct utility function having the generalized Fechner-Thurstone form:

$$(2) \quad U(X; \theta) = \prod_{i=1}^{n} X_i^{\theta_i}; \qquad X_i > 0, \quad i = 1, 2, \ldots, n.$$

For the time being consider the special proper subclass of functions (2) for which the parameters $\theta_i$ have the specified form

$$(3) \quad \theta_i = \beta_i p_i^{\sigma}; \; \beta_i > 0; \qquad \sum_{i=1}^{n} \beta_i = 1; \; \sigma < 1, \sigma \neq 0.$$

In this special case, the parameters $\beta_i$ and $\sigma$ are independent of prices $p_j$ and total expenditure $M$. Now consider the direct utility function

$$(4) \quad V(X) = \left[ \sum_{i=1}^{n} \beta_i^{\frac{1}{1-\sigma}} X_i^{\frac{\sigma}{\sigma-1}} \right]^{-1/\sigma},$$

defined on the same domain as (2), $\beta_i$ and $\sigma$ having the same magnitudes as in (3).

When maximized subject to the usual linear budget constraint, the utility functions (2), (3), and (4) yield identically the same system of demand equations, namely,

$$(5) \qquad X_i = \beta_i p_i^{\sigma-1} M \Big/ \sum_{k=1}^{n} \beta_k p_k^{\sigma}; \; i = 1, 2, \ldots, n.$$

---

*Texas A&M University. Our research was supported by National Science Foundation grant SOC-77-27068. We are indebted to R. C. Battalio, R. M. Harstad, Dale W. Jorgenson, and Stephen White, and the unknown referees for helpful comments and criticisms. We bear sole responsibility for errors and omissions.

[1] Paul Samuelson (1947, pp. 118–21) and others, notably Peter Kalman (1968), and Pollak (1977, 1978) have studies some of the formal properties of individual consumers' direct utility functions when the parameters of the latter are dependent on the commodity prices appearing in the budget constraint.

The system (5) possesses the properties of zero homogeneity, symmetry, and semidefinite negativity (see Basmann, 1968, pp. 159–63).

Let $R_i^{(k)}(X)$ designate the marginal rate of substitution of commodity $k$ for commodity $i$ as determined by the price-dependent direct utility function (2)–(3). Since

$$(6)\qquad R_i^{(k)}(X) = \frac{X_k \beta_i}{X_i \beta_k}\left(\frac{p_i}{p_k}\right)^{\sigma};\; i, k = 1, 2, \ldots, n,$$

the marginal rate of substitution of $R_i^{(k)}$ is in general dependent on the price ratio $p_i/p_k$ at each point $X$ in the domain. Consequently, $U(X;\theta)$ cannot validly be related to the fixed preference direct utility function, $V(X)$, defined by (4), by the formula (1).

Under the conditions of any revealed preference experiment, for example, such as R. C. Battalio et al. (1973), it is in principle impossible to distinguish between the alternative direct utility functions (2), (3), and (4). This observational equivalence of fixed preference direct utility functions and price-dependent utility functions belonging to the generalized Fechner-Thurstone class (2) is by no means a mere curiosity that is limited to the special explicit fixed preference utility function (4). Let $X^*(p, M)$ designate a system of demand functions $X_i^*(p, M)$ (estimated from a comprehensive and faultlessly conducted revealed preference experiment) and possessing the properties of zero degree homogeneity, symmetry, and semidefinite negativity. Suppose there exists a fixed preference direct utility function $V^*(X)$ which, when maximized subject to $p \cdot X = M$, yields the system $X^*(p, M)$. Then the class of all generalized Fechner-Thurstone direct utility functions defined by (2) contains a definite member $U(X;\theta^*)$ with price-dependent parameter $\theta^*$, namely,

$$(7)\qquad U(X;\theta^*) = \prod_{i=1}^{n} X_i^{\theta_i^*};\; X_i > 0,$$

$$\theta_i^* = p_i X_i^*(p, M).$$

When (7) is maximized subject to $p \cdot X = M$, it yields the same system of demand functions $X^*(p, M)$ as the fixed preference utility function $V^*(X)$. It is easy to show that, in general, $U(X;\theta^*)$ and $V^*(X)$ are not related by an equation (1).

Consequently, (7) and $V^*(X)$ are observationally equivalent with respect to the empirical outcomes of any revealed preference experiment. Nonetheless, there are other experimental conditions under which outcomes can distinguish empirically between (7) and $V^*(X)$. Consequently, they are empirically meaningful alternative hypotheses in Samuelson's sense (1947, p. 121). The empirical distinguishability is made clear with help of the special two-commodity case of (2), (3), and (4) specified by

$$(8)\qquad \beta_1 = \beta_2 = 0.5;\; \sigma = 0.5.$$

Given $p = (p_1^0, p_2^0)$ and $M^0$, a definite isoquant (level curve) of

$$(9)\qquad u(\mathbf{X};\theta) = X_1^{0.5\sqrt{p_1}} X_2^{0.5\sqrt{p_2}}$$

and a definite isoquant of

$$(10)\qquad V(X) = \left(X_1^{-1} + X_2^{-1}\right)^{-2},\qquad X_i > 0,\; i = 1, 2,$$

are tangent to the budget line $p^0 X^0 = M^0$ at the same point, $X^0$, but they do not coincide elsewhere. Suppose we want to test whether the consumer's indifference map is given by the price-dependent utility function (9) against the alternative that the consumer's indifference map is given by the fixed preference utility function (10). Battalio and R. M. Harstad have suggested (in private discussion with us) the following incentive-compatible experiment as the briefest illustration that makes our point adequately.[2] Let the market prices $p_1^0$, $p_2^0$ be fixed as well as total expenditure, $M^0$. The consumer purchases

---

[2] Obviously, this experiment is not intended to distinguish between some utility function $V(X)$ and the $W(X;p)$ defined by (1).

$X_1^0, X_2^0$. Keeping $p_1^0, p_2^0, M^0$ fixed, the experimenter offers to take the fixed amount $X_2^0 - X_2^1$ of commodity 2 from the consumer in return for an amount of $\Delta X_1 = X_1^1 - X_1^0 > 0$ to be determined in the following way. The experimenter draws at random a positive number $\Delta X_1^e$ (say) from a suitable range of numbers written on slips of paper. The consumer is instructed to say how much, $\Delta X_1^c$, additional commodity 1 he requires to be exactly as well-off as at $X^0$. If $\Delta X_1^e \geqq \Delta X_1^c$, then the consumer receives $\Delta X_1^e$, that is, he receives more than (he says) he requires; if $\Delta X_1^e < \Delta X_1^c$, then by design of experiment, no trade takes place. Under these conditions, the consumer has an incentive to reveal where his actual indifference curve through $X^0$ intersects the ordinate $X_2^1$. The experiment has to be repeated over a wide range of starting points. It may turn out after many replications, of course, that the hypothesis (9) would be rejected in favor of (10), that is, in favor of the traditional hypothesis that the consumer's indifference map is independent of prices.

## REFERENCES

**Basmann, R. L.**, "Hypothesis Formulation in Quantitative Economics: A Contribution to Demand Analysis," in *Papers in Quantitative Economics*, Lawrence: University Press of Kansas, 1968.

**Battalio, R. C. et al.**, "A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases," *Western Economic Journal*, December 1973, *11*, 411–28.

**Kalman, Peter J.**, "Theory of Consumer Behavior When Prices Enter the Utility Function," *Econometrica*, October 1968, *36*, 497–510.

**Pollak, Robert A.**, "Price-Dependent Preferences," *American Economic Review*, March 1977, *67*, 64–75.

______, "Endogenous Tastes in Demand and Welfare Analysis," *American Economic Review*, June 1978, *68*, 374–79.

**Samuelson, Paul A.**, *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

# Stabilization Policy and Endogenous Wage Stickiness

*By* JOHN MCCALLUM*

Stanley Fischer (1977) and Martin Baily (1978) have developed cases for stabilization policy under rational expectations, but both are open to criticism for their assumptions regarding exogenous contracts of a kind "which seem to be in no one's self-interest" (Robert Lucas, 1981, p. 564). As observed by Brian Kantor, "How the contractual arrangements would alter if the authorities attempted to make use of these apparent rigidities is also not indicated" (1979, p. 1437). The first purpose of this paper is to place the Fischer and Baily arguments in a context of endogenous wage stickiness and to demonstrate that the case for stabilization policy is strengthened, rather than weakened, when this is done. The second purpose, in response to Lucas, is to offer suggestions as to why wage contracts are often written in terms of nominal variables. A more general theme that emerges from the discussion is that the stabilization function might best be treated as a joint venture between the private and public sectors, and that the optimal mix of public and private stabilizing "instruments" is likely to vary considerably across countries. This analysis suggests that the potential gains from countercyclical policy are likely to be greater in the United States than in most other countries.

### I. Supply and Demand for Sticky Wages

Private agents are assumed to choose the optimum degree of wage stickiness, balancing the benefits against the costs. For the representative industry or firm, the degree of wage stickiness may be measured by the length of implicit or explicit wage contracts, and, for the moment, the discussion is limited to the case of unindexed contracts which allow firms to maximize profit subject to a predetermined money wage. The costs of long-term contracts take the form of relatively large deviations from market-clearing output in the event of unanticipated disturbances.[1] Benefits may arise for several reasons, and for the purposes of this section it is assumed, following Jo Anna Gray (1978), that there is a fixed cost in negotiating a contract, so that the benefit from longer contracts consists of a lower negotiating cost per period.

This situation is portrayed in Figure 1, where $\sigma$ is a measure of output instability (or the dispersion of output about its market-clearing level) and $n$ is the length of the wage contract. Units are defined such that each unit of instability represents a one dollar economic cost. Instability rises with contract length as long as forecasts of the near future tend to be more accurate than forecasts of the more distant future. Thus, the $TC$ curves are upward sloping, and for simplicity they are drawn as straight lines. The positive vertical intercept reflects the Fischer-Gray assumption that money wages must be set at least one period before complete information becomes available. The curve $TC_0$ applies to the case of a fixed money supply rule, and the corresponding marginal cost curve $MC_0$

---

*Université du Québec a Montréal and Simon Fraser University.

[1] Robert Hall and David Lilien (1979) show that standard wage contracts may be efficient in the sense that the predetermined relation between remuneration and labor input corresponds to the slope of the labor supply curve. However, such arrangements do not affect the instability arising from aggregate disturbances, since "Within the framework of employment-contingent bargains, there is no way to provide employers with a separate incentive to stabilize employment in the face of an economy-wide reduction in demand" (p. 877). Robert Barro (1977) considers an arrangement where the wage is fixed by contract but employment is maintained at its market-clearing level with the help of appropriate "side payments" between workers and firms. As argued by Fischer (1977b), this kind of contract appears not to exist, perhaps for reasons relating to moral hazard, bargaining costs, monitoring costs, and the costs of specifying future contingencies.

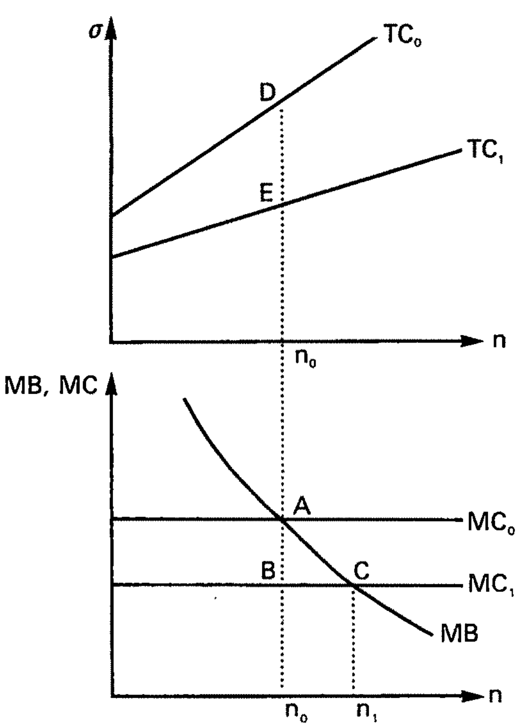


FIGURE 1

is drawn in the lower half of the diagram. The marginal benefit curve ($MB$) in the diagram reflects the assumption that the negotiating cost per period declines as contract length rises. (If $C$ is the fixed negotiating cost, then the cost per period is $C/n$ and $MB$ is equal to $C/n^2$.) In the initial position, under a constant money supply rule, the optimum point is at $A$.

Fischer showed that when the contract length exceeds the reaction time of the monetary authorities, there is a role for stabilization policy because the authorities are able to take account of information that was not available when some contracts were signed. The effect of the policy is to remove or reduce the persistent component of aggregate disturbances, and for this reason, as Baily has shown formally, the accelerator mechanism is dampened. In terms of the diagram, both the Fischer and Baily effects lead to a lower value of $\sigma$ corresponding to any given contract length $n$. The result is a downward shift in both the intercept and the slope of the $TC$ curve, and hence a downward shift in the $MC$ curve as well.[2] However, because of local disturbances, the marginal cost of increasing the length of contracts remains positive. Although both Fischer and Baily assume a fixed contract length, it can be seen from Figure 1 that the effect of the policy change is to increase the optimal contract length from $n_0$ to $n_1$. As a part of the stabilizing function is shifted from the money wage to the money supply, the private sector enjoys part of the welfare gain in the form of increased output stability (an income effect), and part of it in the form of gains arising from longer contracts or career labor markets (a substitution effect). In terms of Figure 1, the income effect is given by $DE$ and the substitution effect by $ABC$.

It appears that the postwar experience of the United States is at least consistent with these predictions. In the late 1950's, there was a rather sudden shift to wage contracts subject to predetermined annual wage increases, with such contracts making up less than 20 percent of all major agreements in 1954 and about 70 percent in 1958 (Joseph Garbarino, 1962, pp. 77–78). It is possible that *ex ante* perceptions of stabilization policy were a factor in this shift, whether or not this early optimism was borne out by later events.[3] Also, it is well known that

---

[2] More formally, the $TC$ curve resulting from a Fischer-Gray model takes a form such as

$$\sigma = B + \left[m_A(\beta+\lambda)^2\sigma_A + m_L\sigma_L\right]f(n),$$

where $\sigma$ is the variance of output about its market-clearing level, $B$ is a constant, $\sigma_A$ and $\sigma_L$ are the variances of aggregate and local disturbance terms, $m_A$ and $m_L$ are related to aggregate and local "multipliers" reflecting the impact of the corresponding disturbance on actual-minus-equilibrium output, $\lambda$ is a constant, and $\beta$ measures the extent of feedback in the monetary rule ($\beta = 0$ implies a constant money supply rule). Finally, $f(n)$ is a function of contract length such that $f(1) = 0$ and $f' > 0$. The constant term $B$ reflects the one-period lag in information and is equal to $m_A\lambda^2\sigma_A + m_L\sigma_L$. The Fischer effect results from setting $\beta = -\lambda$, while the Baily effect takes the form of an induced reduction in $m_A$.

[3] According to Garbarino: "By 1955, the American economy had experienced ten years of fairly high level postwar prosperity and had weathered two minor recessions. The basis for concluding that a new economic era based on government's long term commitment to stability and on industry's rationalized long range planning was at hand" (p. 80).

output has been much more stable in the postwar period than in the earlier decades of the twentieth century (Alan Blinder, 1979, p. 213). In the light of the Lucas critique, direct tests of the effects of stabilization policy based on econometric simulations seem to be of limited value, but it can at least be said that the observed increase in the quantity of wage stickiness and the observed reduction in its "price" are consistent with the present analysis.

## II. Behind the Supply and Demand Curves

> These models.... leave some scope for effective, activist monetary policy due to the fact that people are locked into a nominally fixed arrangement. None of these models offers an explanation as to why people should choose to bind themselves to contracts which seem to be in no one's self-interest.
>
> [Lucas, p. 564]

There are two aspects to this statement. First, why do people lock themselves into arrangements? Second, why are these arrangements nominally fixed? In terms of the supply and demand curves of Figure 1, the first question relates to the demand for wage contracts, and the second asks why wage contracts are not indexed in such a way as to replicate the effects of the optimum monetary rule. If such indexing formulas were in universal use, then there would be no aggregate instability costs arising from long contracts, and hence no gains from stabilization policy. One could take the position that these points are irrelevant to the case for activist policy developed above. People do lock themselves into nominally fixed arrangements, and they do pay a price for such arrangements. By their behavior, people reveal the existence of benefits, and this is sufficient to make the case for policy activism, whether or not the nature of these benefits is well understood by economists. Despite this possible argument, the purpose of this section is to offer a more illuminating response to the points raised by Lucas.

There are several reasons why people may choose to lock themselves into arrangements. Considerations relating to bargaining costs, specific human capital, and the dangers of hold-up costs and opportunistic behavior call for rules or contracts that are infrequently changed but not necessarily nominally fixed (Benjamin Klein et al., 1978; Oliver Williamson et al., 1975). Employee risk aversion would imply a desire to avoid frequent changes in the terms of the "insurance policy" offered by the firm, but here there would be a preference for real wage stability rather than nominal wage stability. Layoffs and stable wages may be preferred to wage flexibility because of effects of wage cuts on employee morale or effort (Robert Solow, 1979), because "the best workers would quit" if wages were cut (Andrew Weiss, 1980), or because wage flexibility may impose monitoring cuts on workers (John Chant, 1981). However, none of these arguments would seem to rule out indexing formulas based on aggregate indicators such as the Consumer Price Index. Even a Keynesian concern with relative wages would be fully met if all money wages were indexed to the same variable, or if the wages paid by a particular firm were indexed to the general level of money wages. In the light of these considerations it will be assumed that people avoid recontracting or frequent changes in wage-setting rules, but, up to this point, no argument has been advanced in favor of nominal wage rigidities or against the use of indexing.

Real world indexing formulas typically index midcontract wages wholly or partially to the Consumer Price Index. Such formulas operate with a lag, if only because of the delay in the publication of official statistics, and in practice, except under conditions of hyperinflation, the lag is unlikely to be less than two or three months. There are at least three reasons why these formulas are a less efficient stabilizer than a Fischer-type monetary rule, which concern the degree of indexing, the universality of indexing, and the fact that indexing is lagged. It is well known that less than complete indexing is desirable in a world that is subject to both demand shocks and supply shocks. The choice of the degree of indexing presumably reflects an assessment as to the probable relative importance of the two kinds of shocks over the life of the contract, where that assessment must be made at the time the contract is signed (Gray, 1976). Such assessments, however, will often be wrong: even if there is a constant money supply rule, there may be unpredictable shifts in the demand for money, and it is certainly difficult to predict supply shocks such as changes in oil prices, crop failures, and changes in productivity growth. By contrast, the monetary authority has an informational advantage because it is not obliged to make such long range predictions.

Second, even if there are no supply shocks, full indexing will not necessarily be desirable unless all other contracts are also fully indexed. A firm selling at a nominally fixed price will not wish to pay full indexed wages, while a firm purchasing inputs at a nominally fixed price will not wish to fully index the price of its product (Robert Gordon, 1982). Workers, too, may oppose indexing of their wages if the Consumer Price Index does not provide a good measure of their cost of living, if the tax system is unindexed, or if they are concerned about their wages relative to those of another group that is not subject to indexation. These considerations suggest that indexing may develop gradually and spread only partially, particularly if there are major costs or institutional barriers to indexation in some sectors of the economy. Examples of such sectors might include imported inputs (Rudiger Dornbusch, 1982), financial instruments, and the tax system. Third, even under universal indexing without supply shocks, the lag in real world indexing rules makes them inferior to a monetary rule. Because of this lag, adjustment to shocks under an indexing rule is analogous to the textbook case of adjustment under adaptive expectations. By contrast, under the monetary rule, money and output adjust in finite time rather than asymptotically to disturbances, implying a shorter period of disequilibrium and hence a more stable economy.

The indexing rule equivalent to the monetary rule must be "anticipatory" rather than "adaptive" with respect to the price level, it must be universally applied, and it must incorporate the latest information on the probable real-nominal mix of future shocks. Given a complete model of the economy, it is possible to develop an expression for such a rule in terms of observable variables, but the rule would be extremely complicated, as it would in general be a function of many variables and depend on all the parameters of the system. Even assuming that such a complicated rule could be universally applied, it would still in all likelihood be inferior to the corresponding monetary rule. This is because the indexing formula, unlike the monetary rule, would have to be based on the structure of the economy $n$ periods in advance of the current period. Unpredictable structural changes such as changes in financial institutions or energy prices may alter parameter values, while information regarding the probable degree of persistence of a particular shock is likely to be available only at or near the time of the shock. Indexing could be truly equivalent to stabilization policy only if the universally applied indexing rule was itself subject to continuous revision, but this would be tantamount to abandoning contracts altogether.

This last point indicates why, even under the most favorable conditions, preset indexing rules could not be a perfect substitute for the corresponding monetary rule. More important from a practical point of view are the first two points relating to nonuniversality and supply shocks. It is these two factors which would seem to account for the very partial degree of indexing in the United States, and to explain why, for lack of a better alternative, people choose to lock themselves into nominally fixed arrangements. It should be noted, however, that this conclusion does not apply to all countries. Some countries have achieved high degrees of universality and standardization through the use of centrally determined indexing rules, and, in some cases, governments or their agencies have been able to suspend the standard indexing rule in the light of special circumstances. In Holland, for example, the government suspended indexation during both of the two oil price shocks of the 1970's while in Australia, until very recently, a central agency set the degree of indexing that would apply across the country on a quarterly or semiannual basis. In these cases, indexation may exert a more powerful stabilizing influence than it does in the United States, but it is clear that such systems involve a greater degree of government intervention than would be required to operate a countercyclical monetary or fiscal policy.

### III. Public and Private Stabilization

> ...[T]he key assumption of Keynesian analysis is the inefficiency of some aspects of private sector activity in comparison to corresponding activities carried out by the government.
> [Barro, 1979, p. 54]

> To suppose that a flexible wage policy is a right and proper adjunct of a system which on the whole is one of *laissez-faire*, is the opposite of the truth. It is only in a highly authoritarian society, where sudden, substantial, all-round changes could be decreed that a flexible wage-policy could function with success.
> [John Maynard Keynes, 1936, p. 269]

At one level, the response to Barro's statement is to be found in *The General Theory*, ch. 19. It is possible to offset aggregate disturbances with changes in the aggregate level of money wages, and in this case the "corresponding activity carried out by the government" is to offset the disturbances with changes in the quantity of money. If money wage flexibility is more costly or more difficult than money supply flexibility, then, indeed, this aspect of private sector behavior is inefficient in comparison with the alternative provided by the government. Also, as Peter Howitt (1981) points out, it may be costly for the private sector to use or process macroeconomic information even if that information is provided free of charge by the government.

More generally, however, the approach of this paper suggests that the stabilizing function might better be treated as a joint venture between the public sector and the private sector rather than as the exclusive domain of either sector alone. A countercyclical policy cannot be a perfect substitute for wage flexibility, if only because stabilization policy is unable to offset local disturbances; but such a policy may nevertheless contribute to stability and reduce the degree to which the burden of stabilization is borne by the money wage. Private indexing may also contribute to stability under certain conditions, but it has been shown that private indexing may be a more effective stabilizer if it is combined with a degree of central direction or coordination. A key assumption implicit in this joint venture approach is that the government and the market are aiming at the same target; that is, the market-clearing level of output. This will not be the case if policy is subject to a political business cycle and/or the market is afflicted by real wage resistance; but the issue addressed in the recent theoretical literature has been whether or not there is a potential role for stabilization policy, under competitive conditions, and not whether policy as actually conducted has been successful.[4]

The optimum mix of private and public stabilizing instruments is likely to vary across countries for several reasons, including differences in historical conditions and differences in the extent of industrial conflict. On grounds of containing industrial conflict or reducing the costs of strikes, it may be desirable to have long wage contracts in the United States (Gordon) and 100 percent indexing in Australia (Noel Gaston, 1981; Anne Romanis Braun, 1976), whereas from a narrowly defined stabilization point of view, short contracts and partial indexing are likely to be preferred. Synchronized or centralized wage bargaining may contribute to stability in countries with a corporatist tradition where unions regard themselves as social partners, but such a system may be destabilizing in countries lacking these conditions. The United States, with its decentralized, noncorporatist traditions and with its relatively high level of industrial conflict, would seem to lack the conditions favoring both short contracts and centralized wage bargaining or indexing schemes. Therefore, because it has relatively few stabilizing instruments at its disposal, the United States may have more to gain from countercyclical policy than most other countries.

---

[4] It could be argued that because of the political business cycle, it is not in the interest of policymakers to aim to stabilize the economy at its market-clearing level of output. Assar Lindbeck, for example, speaks of stabilization policy when politicians are endogenous. If the costs arising from such political factors outweigh the benefits described in this paper, then a case could be made for abandoning stabilization policy altogether. A better, if more difficult, solution may be to devise rules or institutions designed to smooth out both the political business cycle and the regular business cycle.

## REFERENCES

**Baily, Martin N.**, "Stabilization Policy and Private Economic Behavior," *Brookings Papers on Economic Activity*, 1:1978, 11–50.

**Barro, Robert J.**, "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

______, "Second Thoughts on Keynesian Economics," *American Economic Review Proceedings*, May 1979, *69*, 54–59.

**Blinder, Alan**, *Economic Policy and the Great Stagflation*, New York: Academic Press, 1979.

**Brainard, William**, "Uncertainty and the Effectiveness of Policy," *American Economic Review Proceedings*, May 1967, *57*, 411–25.

**Braun, Anne Romanis**, "Indexation of Wages and Salaries in Developed Economies," *IMF Staff Papers*, March 1976, *23*, 226–71.

**Chant, John**, "Lay-Offs and the Enforcement of Employment Contracts," Discussion Paper 81-06-01, Simon Fraser University, 1981.

**Dornbusch, Rudiger**, "PPP Exchange-Rate Rules and Macroeconomic Stability," *Journal of Political Economy*, February 1982, *90*, 158–65.

**Fischer, Stanley**, (1977a) "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

______, (1977b) "Long-Term Contracting, Sticky Prices, and Monetary Policy: A Comment," *Journal of Monetary Economics*, July 1977, *3*, 317–23.

**Garbarino, Joseph W.**, *Wage Policy and Long-Term Contracts*, Washington: The Brookings Institution, 1962.

**Gaston, Noel**, "The Australian Experience with Wage Indexation: 1975–80," mimeo., 1981.

**Gordon, Robert J.**, "Why U.S. Wage and Employment Behavior Differs from that in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

**Gray, Jo Anna**, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

______, "On Indexation and Contract Length," *Journal of Political Economy*, February 1978, *86*, 1–18.

**Hall, Robert E., and Lilien, David**, "Efficient Wage Bargains under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

**Howitt, Peter**, "Activist Monetary Policy under Rational Expectations," *Journal of Political Economy*, April 1981, *89*, 249–69.

**Kantor, Brian**, "Rational Expectations and Economic Thought," *Journal of Economic Literature*, December 1979, *17*, 1422–41.

**Keynes, John Maynard**, *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.

**Klein, Benjamin et al.**, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

**Lindbeck, Assar**, "Stabilization Policy in Open Economies with Endogenous Politicians," *American Economic Review Proceedings*, May 1976, *66*, 1–19.

**Lucas, Robert E., Jr.**, "Tobin and Monetarism: A Review Article," *Journal of Economic Literature*, June 1981, *19*, 558–67.

**Solow, Robert M.**, "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics*, Winter 1979, *1*, 79–82.

**Weiss, Andrew**, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

**Williamson, Oliver E. et al.**, "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange," *Bell Journal of Economics*, Spring 1975, *6*, 250–78.

# Liquidity Preference as Behavior Toward Risk is a Demand for Short-Term Securities—Not Money

*By* WINSTON W. CHANG, DANIEL HAMBERG, AND JUNICHI HIRATA*

In 1958, James Tobin generalized the Keynesian theory of liquidity preference by means of his famous portfolio model in which the demand for money (narrowly defined) is treated as "behavior towards [interest] risk." Whatever merit this theory may have had then has long since been questionable. The reason is the existence of a large set of substitutes for money, typically short-term money market instruments, which can be regarded as riskless, or virtually so, and which pay substantial interest. The availability of these instruments would appear to make Tobin's theory that money is held to cope with interest risk resemble a scenario without a recognizable cast of actors.

In the literature on monetary theory, other authors have also expressed misgivings about the Tobin theory by noting that savings and (nontransferable) time deposits have the same risk properties as money but pay interest (see, for example, Robert Barro and Stanley Fischer, 1976). Although correct, the allusion to these deposits is simplistic. It is true that while both types mimic money's freedom from interest risk in the conventional sense of capital loss, time deposits are still exposed to a kind of interest risk, because they can be liquidated before maturity only with interest penalty. More important, business firms are either denied access to savings deposits or, as in the United States, can hold a maximum of $150,000 (per account) at commercial banks, thereby effectively eliminating large firms as holders. Furthermore, although business firms can own most time deposits, they typically do not (except for negotiable CDs, a money market instrument); they are loath to tie up funds in long-term maturities, and they can usually obtain the same or higher yields on other types of short-term debt instruments that are also negotiable. In the United States, households have long accounted for about one-third of demand deposits, business firms owning most of the rest. Therefore any effort to rest a case against the Tobin theory of money demand on the existence of savings and time deposits gets at only a small part of the problem. This stricture extends to so-called NOW and ATS accounts. These interest-bearing demand deposits (disguised under other names) are also denied to business firms. For them, the short-term instruments of the money market are the principal alternative to money in asset portfolios.

## I. Money Market Instruments

These instruments have received limited attention in the money literature. Therefore, we must describe their various characteristics that highlight their closeness to money with respect to freedom from interest risk.

Some of the money market instruments are literally riskless. Repurchase agreements (*RP*s) between banks (as well as securities dealers and large savings and loan associations) and nonbank investors are a notable example. They are devoid of interest risk because the borrower must repurchase the securities involved in the loan at their *original* price plus interest. Moreover, over half of the approximately $169 billion of *RP*s recently outstanding in the United States were either overnight in term or callable at any time.[1] Even when maturity dates are fixed, beyond one day, investors unexpectedly needing cash can form *reverse RP*s with the collateral they have taken in, thereby avoiding liquidation risk.

---

*Chang and Hamberg: professors of economics, State University of New York–Buffalo; Hirata: assistant professor of economics, Technological University of Nagaoka, Japan.

[1] These and other numbers on money market instruments are based on data in the *Federal Reserve Bulletin*, July 1982, Tables 1.24, 1.29, 1.32, 1.42, 1.44, and 3.14; and Federal Reserve Bank of St. Louis, *Monetary Trends*, June 24, 1982, p. 6.

Equally liquid are federal-funds loans to commercial banks by other financial institutions; these are mostly overnight loans with specified interest rates.

Other money market instruments, such as Treasury bills, CDs, and bankers' acceptances, do possess positive, if limited, interest risk when sold before maturity.[2] However, these instruments have highly developed secondary markets in which they can be purchased with as short maturities as investors desire and thus are virtually riskless. Also, since 1975, markets for interest rate futures contracts have been developed that permit investors to hedge against (or manage) interest risk on Treasury bills. Other instruments, such as commercial paper, have limited secondary markets, but their maturity dates are (like those on *RP*s and CDs) fixed to coincide with tax and dividend, payroll, and any other cash-need, dates of investors. Moreover, most issuers of commercial paper will normally redeem this paper before maturity from hard-pressed investors. And interest rate futures also exist for commercial paper. The total of all money market instruments extant in the United States borders on $1 trillion.

Mention should be made, too, of accounts in money market (mutual) funds. The assets of these funds consist entirely of the whole array of money market instruments, except for a tiny amount of demand deposits. If anything, the accounts in these funds are even more liquid than the instruments themselves, because checks may be written upon them. These funds aggregate to some $200 billion. They appeal not only to households, but to small business firms and small institutional investors generally; all are too small to justify the cost of a money market specialist to manage their portfolios.

As noted, the existence in large magnitudes of such a rich assortment of liquid, money market instruments that investors can deploy to suit their various needs, with all devoid of interest risk or nearly so and all paying abundant interest, suggests the extreme dubiety of a theory that purports to explain money demand as a response to interest risk. This skepticism is buttressed by the frantic pace of development of cash-management techniques used by business firms (see Richard Porter, Thomas Simpson, and Eileen Mauskopf, 1979). In general, these techniques are designed to reduce perceived variance of cash flows and thus enable investors to reduce their cash holdings and increase the amounts available for investment in money market instruments.

## II. A Three-Asset Model

Doubt and skepticism are not enough. They should be supported by argument at least as rigorous as that of the original (Tobin) argument. To this end, we take up the "risk" demand for money in a three-asset model that contains money, a virtually riskless asset (representing the foregoing spectrum of money market instruments), and a risky asset. The model is essentially an extension of Tobin's (1958) two-asset (money-consol) formulation. Our purpose is to show that if money market instruments are explicitly incorporated in a portfolio model, the demand for money as a buffer against interest risk vanishes. A model with at least three assets of the types described must be formulated, because some of the (nonmonetary) assets do possess significant interest risk, even while others have little or none, but both types are relevant to the portfolio decision.

---

[2] Money market instruments are of two types: interest-bearing instruments or securities and discount paper. The former are exemplified by large (negotiable) certificates of deposit (CDs), *RP*s, and municipal notes; all pay interest at maturity. Discount paper consists of Treasury bills, commercial paper, and bankers' acceptances. None bears interest; instead, interest is earned on them because they are issued at a discount from face value. For such instruments with a current maturity of six months or less, so that there is no compounding, the equivalent bond yield $r$ is given by the formula $r = (365/m)[(F-P)/P]$, where $m$ is the number of days to maturity, $F$ the face value, and $P$ the market price. Solving this equation for $P$ gives $P = F/[1+(rm/365)]$, which can also be used for interest-bearing instruments. The elasticity $e$ of $P$ with respect to $r$, which we can take as a measure of liquidity, is $e = -1/[1+(365/rm)]$. From this it can be seen that $e \rightarrow 0$ as $m \rightarrow 0$. Apart from overnight *RP*s, the current maturities of most money market instruments lie within a range of 1 to 91 days. At interest rates of 15 and 10 percent, for example, for maturities of 91 days, $e = -3.6$ percent and $-2.4$ percent, respectively; for maturities of 7 days, and the same interest rates, $e = -0.3$ percent and $-0.2$ percent, respectively. These figures for $e$ illustrate the meaning of "limited" interest risk. By comparison, a consol has the highest price elasticity of all: $-100$ percent.

Utilizing Tobin's formulation signifies the application of Harry Markowitz's (1952) mean-variance model to the analysis of money demand. Although this model is widely accepted, it has some limitations. Indifference curves between mean and variance can be drawn only if an investor's preferences can be expressed by a quadratic utility function, or the probability distributions of returns on portfolios are normal (see, for example, Tobin, 1969). In view of strong criticism of the quadratic utility function in the present context (for example, Kenneth Arrow, 1970), we have adopted the normal distribution in our analysis.[3]

Our three-asset model is set up in the framework of a nonlinear program. This procedure is dictated by our interest in the conditions and properties of a corner solution marked by zero money holdings. In the absence of the quadratic utility function, results will be derived without explicit specification of a utility function. Then asset-demand functions will be obtained for two (of seven possible) solutions: the corner case just referred to and the interior solution. By way of a specific example, demand equations will be derived for the class of utility functions corresponding to constant absolute risk aversion. This will be followed by a sketch of same for a quadratic utility function.

### A. *Assumptions*

As remarked, our model basically extends that of Tobin's (1958) by introducing a third, virtually riskless, asset. For convenience, we label our three assets money (narrowly defined), a short-term asset, and a long-term one. Investor behavior is the outcome of an expected utility maximization for the choice of assets.

The interest rate for each asset $r_i$ $(i=1,2,3)$ is assumed to have the following properties: $r_1=0$, $r_2>0$, and $r_3>0$.[4] Denoting the capital gain or loss of each asset by the random variables $g_i$ $(i=1,2,3)$, we assume that money has no capital gain or loss, that is, $g_1=0$. The investor's estimated distribution function for capital gains or losses has a joint normal distribution; the means of the $g_i$ are zero, the variances of $g_2$ and $g_3$ are $v_2$ and $v_3$, and their covariance is $c=pv_2^{.5}v_3^{.5}$, where $p$ is the correlation coefficient between $g_2$ and $g_3$. It is assumed that $-1<p<1$. Thus the measure of return on each asset is the expected value of its interest rate, and the measure of asset risk is the variance of capital gain or loss. The short-term asset is assumed less risky than the long-term one, that is, $v_2<v_3$. Since $g_1=0$, so does $v_1$.

Following Tobin (1958), we assume the share of each asset $A_i$ to be nonnegative and independent of initial wealth. The wealth constraint implies that these shares add to one.

The return $R$ from a portfolio is

$$R=A_2(r_2+g_2)+A_3(r_3+g_3)$$

where $R$ has a normal distribution with mean $\mu$ and variance $v$.

Now, in the aforementioned framework of a nonlinear programming model, we can specify the maximization problem as[5]

$$(1)\qquad Max\,E[f(R)]=Max\,U(\mu,v),$$

---

[3] There is empirical support for the assumption of normality in the distribution of portfolio returns. See Eugene Fama (1976, pp. 43–44).

[4] Setting $r_1=0$ is done only for convenience and does not affect the substantive conclusions. All that is required is that $r_1$ be the lowest of the three interest rates.

[5] Note that the equality in our objective function need not hold. However, John Chipman (1973) has derived the necessary and sufficient conditions on the utility function to guarantee this equality when the distribution of returns is normal. Hence we shall proceed as if these conditions are satisfied. The utility function characterized by constant absolute risk aversion, employed below to solve the problem explicitly, satisfies the conditions for equality.

subject to

$$(2)\qquad \mu \equiv E(R) = A_2 r_2 + A_3 r_3;$$

$$(3)\qquad v \equiv v(R) = E[R - E(R)]^2$$

$$= A_2^2 v_2 + 2A_2 A_3 c + A_3^2 v_3;$$

$$(4)\qquad A_1 + A_2 + A_3 = 1, \quad A_i \geqq 0 \ (i = 1,2,3).$$

Finally, the utility function $f(R)$ is assumed to have a positive first-order derivative and negative second-order derivative. The investor is thus a risk averter. This assumption implies that the utility function $U(\mu, v)$ has the properties that $U_\mu \equiv \partial U/\partial\mu > 0$ and $U_v \equiv \partial U/\partial v < 0$, and it is concave in $\mu$ for each $v$. It is further assumed that $U_v \neq -\infty$.

### B. *General Solution and its Properties*

In this section, we solve the maximization problem formulated from (1). We begin by deriving, first, the Kuhn-Tucker conditions based on a general utility function.

The Lagrangean function ($L$) is

$$L = U\left(A_2 r_2 + A_3 r_3, A_2^2 v_2 + 2A_2 A_3 c + A_3^2 v_3\right)$$

$$+ \lambda(1 - A_1 - A_2 - A_3).$$

The Kuhn-Tucker conditions are

$$(5)\qquad \partial L/\partial A_1 = -\lambda \leqq 0$$

$$(6)\qquad \partial L/\partial A_2 = U_\mu(\partial\mu/\partial A_2) + U_v(\partial v/\partial A_2) - \lambda \leqq 0$$

$$(7)\qquad \partial L/\partial A_3 = U_\mu(\partial\mu/\partial A_3) + U_v(\partial v/\partial A_3) - \lambda \leqq 0$$

$$(8)\qquad (\partial L/\partial A_1)A_1 + (\partial L/\partial A_2)A_2 + (\partial L/\partial A_3)A_3 = 0$$

$$(9)\qquad \partial L/\partial\lambda = 1 - A_1 - A_2 - A_3 = 0, \quad A_i \geqq 0 \ (i = 1,2,3)$$

$$(10)\qquad \partial\mu/\partial A_2 = r_2, \ \partial\mu/\partial A_3 = r_3$$

$$(11)\qquad \partial v/\partial A_2 = 2(A_2 v_2 + A_3 c),$$

$$\partial v/\partial A_3 = 2(A_3 v_3 + A_2 c).$$

Conditions (5) and (8), plus the nonnegativity of the $A_i$, imply that $(\partial L/\partial A_1)A_1 = -\lambda A_1 = 0$. Therefore, we have the condition (C):

$$(\mathrm{C})\qquad \text{If } \lambda > 0, \text{ then } A_1 = 0;$$

$$\text{if } A_1 > 0, \text{ then } \lambda = 0.$$

Accordingly, we have the following proposition:

PROPOSITION 1: *If* $v_1 = 0$, $v_2 \to 0$, *and* $r_2 > r_1 = 0$, *then* $A_1 = 0$.

PROOF:

Conditions (6), (10), and (11) imply that $\lambda \geqq U_\mu r_2 + 2U_v(A_2 v_2 + A_3 c) > 0$,[6] since $r_2 > 0$, and $v_2 \to 0$ implies $c \to 0$. Therefore, by condition (C), $A_1 = 0$.

The above proposition is valid for the general utility function. One can alternatively prove this point by arguing that if the investor switches out of $A_1$, he is always better off in this case. Setting $v_2 = 0$, we have $U(\mu, v) = U(A_2 r_2 + A_3 r_3, A_3^2 v_3)$. Letting $dA_1 = -dA_2 < 0$ (that is, switching out of $A_1$ into $A_2$ while keeping $A_3$ constant), we obtain $dU = U_\mu r_2 dA_2 > 0$. Thus $A_2$ dominates $A_1$, and the investor will never hold money as a response to interest risk in this case.

[6] By means of the proof here, Proposition 1 can be generalized to the case where $v_2$ does not approach zero. Following the convention used in the finance literature, we define $\beta_2 \equiv (A_2 v_2 + A_3 c)/v$, which is the risk of the short-term asset in the portfolio relative to the risk of the portfolio (see, for example, Fama, p. 60). The condition $U_\mu r_2 + 2U_v(A_2 v_2 + A_3 c) > 0$ can be written as $\beta_2 < (-U_\mu/U_v)(r_2/2v)$, which is a sufficient condition for the domination of money by the short-term asset. Note that if $v_2 \to 0$, then $\beta_2 \to 0$.

The above proposition can be illustrated by the use of the efficiency frontier in the $(\mu, v)$ space. Solving for $A_3$ from (2) and (4) in terms of $A_1$ and $\mu$, substituting the solution into (3), and setting $v_2 = 0$, we obtain the family of opportunity loci by treating $A_1$ as a parameter:

$$\begin{aligned} v &= F(\mu, A_1) \\ &= v_3(\mu - r_2(1-A_1))^2/(r_3-r_2)^2. \end{aligned}$$

Without loss of generality, we assume $r_3 > r_2$. It can be shown that each opportunity locus in the $(\mu, v)$ space is upward sloping and convex downward. Moreover, at constant $\mu$, an increase in $A_1$ increases $v$. A typical opportunity locus for $A_1 = \bar{A}_1$ is drawn in Figure 1 as the $Z_1Z_1'$ curve. When $\bar{A}_1 \rightarrow 0$, the curve shifts to $Z_0Z_0'$, and when $\bar{A}_1 \rightarrow 1$, it shrinks to the origin. At $Z_1$, $A_3 = 0$ and $A_2 = 1 - \bar{A}_1$; at $Z_1'$, $A_2 = 0$ and $A_3 = 1 - \bar{A}_1$. Thus by moving from $Z_1$ to $Z_1'$, $A_3$ is continuously substituted for $A_2$. The $Z_0Z_0'$ locus is the efficiency frontier. At $Z_0$, wealth is held only in asset 2 while at $Z_0'$, only in asset 3.

The optimal choice is determined by maximizing $U(\mu, v)$, subject to the efficiency frontier $Z_0Z_0'$. An optimal solution point is illustrated at $W$ where the indifference curve $U_{max}$ is tangent to the efficiency frontier. The solution for $A_3$ at $W$ is $(v^*/v_3)^{0.5}$ and that for $A_2$ is $1-(v^*/v_3)^{0.5}$.

If the investor is strongly risk averse, the optimal choice can conceivably be at $Z_0$, so that wealth is held only in the virtually riskless asset. This is more likely to be the case as the difference between $r_3$ and $r_2$ narrows. In the case where $r_2 > r_3$, the $ZZ'$ curves become negatively sloped, and the efficiency frontier reduces to a point on the $\mu$ axis. Therefore, the virtually riskless asset will always dominate the risky one.

We conclude that in the presence of a virtually riskless asset with a positive rate of return, the efficiency frontier can be generated only by a set of portfolios with zero money holdings. Liquidity preference as behavior toward risk is thus not a demand for money.

Some further characteristics of the optimal solution are noteworthy. Substituting (5)–(7) into (8), and using (2)–(4), we have

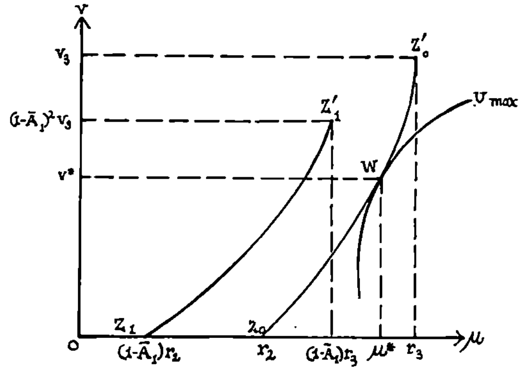


FIGURE 1

$$(12) \qquad \lambda = U_\mu \mu + 2U_v v.$$

From equation (6) and condition (C) it is evident that if $U_\mu(\partial\mu/\partial A_2) + U_v(\partial v/\partial A_2) > 0$ (implying that $\lambda > 0$), then $A_1 = 0$; similarly, from (7) and (C) it follows that if $U_\mu(\partial\mu/\partial A_3) + U_v(\partial v/\partial A_3) > 0$ (implying that $\lambda > 0$), then again $A_1 = 0$. Together, these characteristics yield a set of *sufficient* conditions for an investor to shun $A_1$, as well as a *necessary* condition for holding it:

PROPOSITION 2: *If*
(i) $U_v/U_\mu > -(\partial\mu/\partial A_2)/(\partial v/\partial A_2)$,
*or* (ii) $U_v/U_\mu > -(\partial\mu/\partial A_3)/(\partial v/\partial A_3)$,
*or* (iii) $U_v/U_\mu > -(\mu/2v)$, *then* $A_1 = 0$;
*if* $A_1 > 0$, *then* $U_v/U_\mu = -(\mu/2v)$.

Condition (i) means that a switch from $A_1$ into $A_2$ increases the investor's utility; hence $A_2$ dominates $A_1$; condition (ii) can be interpreted similarly; and condition (iii) means the opportunity cost of holding $A_1$ is positive, implying that $A_1$ should not be held.

If an investor holds all three assets, we have

PROPOSITION 3: *If* $A_1 > 0$, $A_2 > 0$, *and* $A_3 > 0$, *then*

$$\begin{aligned} U_v/U_\mu &= -\mu/2v \\ &= -(\partial\mu/\partial A_2)/(\partial v/\partial A_2) \\ &= -(\partial\mu/\partial A_3)/(\partial v/\partial A_3). \end{aligned}$$

These conditions can be interpreted as meaning that the marginal rate of substitution along a given indifference curve at the optimum point $[U_v/U_\mu = -(d\mu/dv)|_{U=U^*}]$ must be equal to the marginal tradeoff between return and risk of the second and third assets, and must also be equal to the negative one-half of the overall return-to-risk ratio. Alternatively, the conditions imply that, on the margin, if the investor holds all three assets, their marginal utilities $(\partial U/\partial A_i,\ i = 1,2,3)$ must be the same and equal to zero.

### C. *Demand Equations*

In this section, we derive the demand equations associated with the two cases just considered.

*Case* 1: $A_1 > 0, A_2 > 0, A_3 > 0$. From the sign restrictions in equations (5)–(7), together with the nonnegativity of the $A_i$, it follows (in (8)) that this case requires that $\partial L/\partial A_1 (= \lambda) = \partial L/\partial A_2 = \partial L/\partial A_3 = 0$. Thus we have $U_\mu(\partial\mu/\partial A_2) + U_v(\partial v/\partial A_2) = 0$ and $U_\mu(\partial\mu/\partial A_3) + U_v(\partial v/\partial A_3) = 0$, or equivalently:

$$(13)\qquad H\begin{bmatrix} A_2 \\ A_3 \end{bmatrix} = -\frac{U_\mu}{2U_v}\begin{bmatrix} r_2 \\ r_3 \end{bmatrix},$$

where $$H = \begin{bmatrix} v_2 & c \\ c & v_3 \end{bmatrix},$$

the variance-covariance matrix. The solution to (13) is

$$(14)\qquad \begin{bmatrix} A_2 \\ A_3 \end{bmatrix} = -\frac{U_\mu}{2|H|U_v}\begin{bmatrix} v_3 r_2 - c r_3 \\ -c r_2 + v_2 r_3 \end{bmatrix},$$

where $|H| = v_2 v_3 - c^2 = (1-p^2)v_2 v_3 > 0$. In turn,

$$(15)\qquad A_1 = 1 - A_2 - A_3 = 1 - (U_\mu/2|H|U_v) \times [(c - v_3)r_2 + (c - v_2)r_3].$$

*Case* 2: $A_1 = 0, A_2 > 0, A_3 > 0$. Now only equations (6) and (7) are satisfied with strict equality. By eliminating $\lambda$ from them, we obtain

$$(16)\qquad (v_2 - c)A_2 + (c - v_3)A_3 = -(U_\mu/2U_v)(r_2 - r_3).$$

This, together with $A_2 + A_3 = 1$, yields

$$(17)\qquad \begin{bmatrix} A_2 \\ A_3 \end{bmatrix} = \frac{1}{v_2 - 2c + v_3} \times \begin{bmatrix} -(U_\mu/2U_v)(r_2 - r_3) - (c - v_3) \\ (U_\mu/2U_v)(r_2 - r_3) + (v_2 - c) \end{bmatrix},$$

where $v_2 - 2c + v_3 = (v_2^{.5} - v_3^{.5})^2 + 2v_2^{.5}v_3^{.5}(1-p) > 0$.

Let us now utilize these results to derive the demand equations for the utility function corresponding to constant absolute risk aversion, first set forth by Arrow (1970) and John Pratt (1964). In addition to satisfying the Chipman conditions noted in footnote 5, this class of utility functions also satisfies the requirements imposed by Arrow for a utility function appropriate to economic behavior under uncertainty. The function is of the form $f(R) = -exp(-kR)$, $k > 0, f'(R) > 0, f''(R) < 0$, whence

$$(18)\qquad U(\mu, v) = -exp(-k(\mu - kv/2)).$$

In this case, we have

$$(19)\qquad U_\mu/U_v = -2/k.$$

Thus (14) and (15) become

$$(14')\qquad \begin{bmatrix} A_2 \\ A_3 \end{bmatrix} = \frac{1}{k|H|}\begin{bmatrix} v_3 r_2 - c r_3 \\ -c r_2 + v_2 r_3 \end{bmatrix};$$

$$(15')\qquad A_1 = 1 - (1/k|H|) \times [(v_3 - c)r_2 + (v_2 - c)r_3] > 0.$$

In its turn, (17) becomes

$$(17') \quad \begin{bmatrix} A_2 \\ A_3 \end{bmatrix} = \frac{1}{v_2 - 2c + v_3} \times \begin{bmatrix} (1/k)(r_2 - r_3) - (c - v_3) \\ -(1/k)(r_2 - r_3) - (c - v_2) \end{bmatrix}.$$

Some comparative statics analysis is illuminating in this case of constant absolute risk aversion. Our principal interest lies in the effects of a change in $r_2$ and $r_3$ on $A_1$. Therefore, we consider the case where $A_1 > 0$. From (15′) we have

$$(20) \quad \partial A_1/\partial r_2 = (1/k|H|)(c - v_3);$$

$$(21) \quad \partial A_1/\partial r_3 = (1/k|H|)(c - v_2).$$

The money demand schedule is negatively sloped with respect to $r_i$, if and only if $v_j > c$, $i, j = 2, 3$, $i \neq j$. These conditions are clearly satisfied if $g_2$ and $g_3$ are negatively correlated.

Demand equations can also be derived for the quadratic utility function. In this case, the function is of the form $f(R) = (1+b)R + bR^2$ where, for a risk averter, $-1 < b < 0$, and to guarantee $f'(R) = (1+b) + 2bR > 0$, the further restriction that $R < -(1+b)/2b$ is required. In this case,

$$(22) \quad U(\mu, v) = (1+b)\mu + b(\mu^2 + v);$$

$$(23) \quad U_\mu/U_v = [(1+b) + 2b\mu]/b = [(1+b) + 2b(A_2 r_2 + A_3 r_3)]/b.$$

The marginal rate of substitution now involves $A_2$ and $A_3$, making it a bit tedious to obtain explicit solutions for $A_2$ and $A_3$. Equation (23) can be substituted into (14) and (17) to obtain these solutions. This task we leave to the interested reader. An analysis of the $A_1$ demand function can be carried out similar to that just performed for the case of constant absolute risk aversion.

One final remark: Some readers may find our results obvious, especially that of Proposition 1. But, the obvious needs to be shown rigorously; more important, if our results are indeed obvious, readers must explain the approving presence of the Tobin model in so many macro and money texts in print.[7] After all, the relevant substitute assets for money do exist—to the tune of almost \$1 trillion, as noted. Our virtually riskless asset is, therefore, not a fiction of somebody's imagination, so that our results are not fictional. On the contrary, the original Tobin model is.

## REFERENCES

**Arrow, Kenneth J.**, *Essays in the Theory of Risk Bearing*, Amsterdam: North-Holland, 1970.

**Barro, Robert J. and Fischer, Stanley**, "Recent Developments in Monetary Theory," *Journal of Monetary Economics*, April 1976, *2*, 133–67.

**Branson, William H.**, *Macroeconomic Theory and Policy*, 2d ed., New York: Harper and Row, 1979.

**Chipman, John S.**, "The Ordering of Portfolios in Terms of Mean and Variance," *Review of Economic Studies*, April 1973, *40*, 167–90.

**Dernburg, T. F. and McDougall, D. M.**, *Macroeconomics*, 5th ed., New York: McGraw-Hill, 1980.

**Fama, Eugene F.**, *Foundations of Finance: Portfolio Decisions and Security Prices*, New York: Basic Books, 1976.

**Goldfeld, Stephen M. and Chandler, Lester V.**, *The Economics of Money and Banking*, 8th ed., New York: Harper and Row, 1981.

**Laidler, David**, *The Demand for Money: Theories and Evidence*, 2nd ed., New York: Harper & Row, 1977.

**Lovell, Michael C.**, *Macroeconomics: Measurement, Theory and Policy*, New York: John Wiley and Sons, 1975.

**Markowitz, Harry M.**, "Portfolio Selection,"

---

[7] To mention only five: William Branson (1979, pp. 251–58), T. F. Dernburg and D. M. McDougall (1980, pp. 125–27), Stephen Goldfeld and Lester Chandler (1981, pp. 317–22), David Laidler (1977, pp. 88–98), and Michael Lovell (1975, pp. 183–92).

*Journal of Finance*, March 1952, *7*, 77–91.

**Porter, Richard, Simpson, Thomas and Mauskopf, Eileen,** "Financial Innovation and the Monetary Aggregates," *Brookings Papers on Economic Activity*, 1:1979, *9*, 213–29.

**Pratt, John W.,** "Risk Aversion in the Small and the Large," *Econometrica*, January–April 1964, *32*, 122–36.

**Tobin, James,** "Liquidity Preference as Behavior Towards Risk," *Review of Economic Studies*, February 1958, *25*, 65–86.

______, "Comment on Borch and Feldstein," *Review of Economic Studies*, January 1969, *36*, 13–14.

# A Rationale for Preference Reversal

*By* GRAHAM LOOMES AND ROBERT SUGDEN*

Ever since the formulation by John von Neumann and Oskar Morgenstern of a set of axioms of rational choice under uncertainty, a number of situations have been identified in which there are significant and repeated violations of one or more of those basic axioms. A comprehensive survey of relevant evidence has been provided by Paul Schoemaker (1982), together with an outline of various attempts that have been made to explain the observed behavior.

One such attempt was our 1982 article, and was given the name *regret theory*. We showed that several behavioral patterns which contradict standard expected utility theory (the "common consequences effect" or Allais' paradox, the "common ratio effect," and the "isolation effect" in two-stage gambles) are predicted by regret theory; and other patterns, such as the "reflection effect" and simultaneous gambling and insurance, if not firmly predicted, are at least consistent with regret theory.

However, a number of problematical cases were not explicitly treated in our earlier paper. One such case, the phenomenon of preference reversal, was the subject of an article by David Grether and Charles Plott (1979). This phenomenon, first reported by Sarah Lichtenstein and Paul Slovic (1971) and by Harold Lindman (1971), takes the following general form: asked to make a direct choice between gamble $A$ and gamble $B$, an individual states a preference for $A$; but asked to consider the two gambles separately, he places a higher certainty equivalent value, or reservation price, on $B$.

This phenomenon is particularly interesting because, as Grether and Plott pointed out, with such reversals it appeared that "...no optimization principles of any sort lie behind even the simplest of human choices" (p. 623). The purpose of this paper is to suggest that this is not so, and that regret theory may provide the optimization principles which explain the asymmetric pattern of choice observed by Grether and Plott, and others.

## I. An Outline of Regret Theory

The essential notion underlying regret theory is that people tend to compare their actual situations with the ones they would have been in, had they made different choices in the past. If they realize that a different choice would have led to a better outcome, people may experience the painful sensation of regret; if the alternative would have led to a worse outcome, they may experience a pleasurable sensation we call "rejoicing." When faced with new choice situations, people remember their previous experiences and form expectations about the rejoicing and regret that the present alternatives might entail. They then take these expectations into account when making their decisions. We model all this as follows.

In any pairwise choice, individuals are regarded as choosing between pairs of *actions* denoted $A_i$ and $A_k$, where each action is an $n$-tuple of state-contingent *consequences*. The consequence of choosing $A_i$ if the $j$th state occurs is denoted $x_{ij}$; and the probability of the $j$th state occurring is given as $p_j$.

It is proposed that, for any individual, there exists a *choiceless* utility function, $C(.)$, unique up to an increasing linear transformation, which assigns a real-valued utility index number to every conceivable consequence. The value $C(x_{ij})$ is written as $c_{ij}$, and represents the utility an individual would derive if the consequence were experienced other than as a result of choice, for example, if the consequence were imposed in some way.

The choice an individual faces is either to accept $A_i$ and simultaneously reject $A_k$, or else to accept $A_k$ and simultaneously reject $A_i$. Suppose he contemplates accepting $A_i$

---

*Department of economics, University of Newcastle upon Tyne.

and rejecting $A_k$. Should the $j$th state occur, he will experience $x_{ij}$ but simultaneously miss out on $x_{kj}$. So in addition to $c_{ij}$, he may also derive some increment or decrement of utility due to rejoicing or regret: an increment if $c_{ij} > c_{kj}$; a decrement if $c_{ij} < c_{kj}$.

In the simplest restricted form of the model, it is proposed that regret-rejoicing be represented by the function $R(.)$, where the degree of regret-rejoicing depends only on the difference $(c_{ij} - c_{kj})$. We assume that $R(.)$ is strictly increasing and three times differentiable, and that $R(0) = 0$. On this basis, the *modified utility* that results if $A_i$ is chosen and $A_k$ is rejected and the $j$th state occurs can be written as $m_{ij}^k$ where

$$(1) \qquad m_{ij}^k = c_{ij} + R(c_{ij} - c_{kj}).$$

We can then define the expected modified utility of accepting $A_i$ and rejecting $A_k$ as $E_i^k$ where

$$(2) \qquad E_i^k = \sum_{j=1}^{n} p_j m_{ij}^k.$$

The proposed choice rule that the individual will prefer $A_i$, prefer $A_k$, or be indifferent between them according to whether $E_i^k$ is greater than, less than, or equal to $E_k^i$, generates the proposition that

$$(3) \quad A_i \succsim A_k \quad \text{iff} \quad \sum_{j=1}^{n} p_j\big[c_{ij} - c_{kj} + R(c_{ij} - c_{kj}) - R(c_{kj} - c_{ij})\big] \gtreqless 0.$$

For compactness, we define a function $Q(.)$ such that for all $\xi$, $Q(\xi) = \xi + R(\xi) - R(-\xi)$. (Note that for all $\xi$, $Q(\xi) = -Q(-\xi)$.) Thus

$$(4) \quad A_i \succsim A_k \quad \text{iff} \quad \sum_{j=1}^{n} p_j Q(c_{ij} - c_{kj}) \gtreqless 0.$$

If $Q(.)$ is linear, regret theory predicts choices that are consistent with expected utility theory, but if the function is nonlinear, the predictions of the two theories diverge. If $Q(\xi)$ is assumed to be convex for all $\xi > 0$, regret theory predicts a number of patterns of behavior which, although inconsistent with expected utility theory, have been observed in repeated experiments, including those referred to in the introduction to this paper. We shall now show that under the same assumptions, regret theory can explain the main kind of preference reversal observed by Grether and Plott.

## II. Preference Reversal and Regret Theory

Grether and Plott asked individuals to consider pairs of bets where the "*P* bet" offered a high probability of a relatively small money gain, while the "\$ bet" offered a lower probability of a larger gain. Both bets had the same mean value,[1] but the variance was greater for the \$ bet. All pairs of bets can be described with the following notation.

The *P* bet offered the dollar consequences $y_P$ and $x_P$ with probabilities $\Pi_P$ and $1 - \Pi_P$; $y_P > 0 > x_P$. The \$ bet offered the consequences $y_\$$ and $x_\$$ with probabilities $\Pi_\$$ and $1 - \Pi_\$$; $y_\$ > 0 > x_\$$. In all Grether and Plott's experiments, $\Pi_P > 0.5 \geqslant \Pi_\$$ and $y_\$ > y_P$. In some pairs of bets $x_P > x_\$$, while in others $x_P < x_\$$, but in all cases the difference between the two negative consequences was relatively small; in particular:

$$|x_P - x_\$| < |y_P - y_\$|$$

and

$$|x_P - x_\$| < |y_P - x_\$|.$$

Each individual was asked to state the lowest price at which he would sell each of the bets; we shall denote these two reservation prices by $r_P$ and $r_\$$. He was also at a different time asked to make a direct choice between the *P* bet and the \$ bet. In 561 cases (74.4 percent) out of 754,[2] individuals placed

---

[1] In Grether and Plott's experiments, the two bets did not have exactly the same expected money values, but only because consequences were rounded to the nearest dollar and probabilities to the nearest one-thirty-sixth.

[2] These figures are based on Grether and Plott, Tables 5, 6, 8 and 9 (pp. 632–33), but omitting 14 cases where $P \sim \$$, since Grether and Plott gave no information about reservation prices in these cases.

a higher reservation price on the \$ bet, despite the greater variance of those bets. Moreover Grether and Plott reported (p. 632) that the reservation prices for \$ bets were frequently greater than their expected values. These observations can be squared with expected utility theory, but only by dropping the usual assumption of risk aversion.

Much more serious for standard theory, however, was the incidence of preference reversal. In 76 cases (10.1 percent) $P \prec \$$ was combined with $r_P \geqslant r_\$$ and in 205 cases (27.2 percent) $P \succ \$$ was combined with $r_P \leqslant r_\$$.

These observations are completely inconsistent with conventional theory. But now consider the implications of regret theory. Tables 1–3 portray three matrices of state-contingent consequences corresponding to the three choice problems facing an individual for any pair of bets. (The columns in each table represent states of the world, with the probability of each state given above.)

Table 1 represents the choice between retaining the $P$ bet and selling it for the sum $r_P$ (an action we denote by $P^*$). If $r_P$ is the reservation price, then $P \sim P^*$. Similarly in Table 2, $\$^*$ is the action of selling the \$ bet, and if $r_\$$ is the reservation price, then $\$ \sim \$^*$. Table 3 represents the choice between the two bets. This matrix is constructed to correspond with the random device used in Grether and Plott's experiments (see p. 629).

To produce firm predictions, and to highlight the role of regret and rejoicing in our theory, we shall make the additional simplifying assumption that the choiceless utility function $C(.)$ is linear over the relevant range. On that basis, we obtain the following general result.

Let $A_i$ be an action that gives the consequence $w$ with certainty while $A_k$ is an actuarially fair gamble offering $w+u$ with probability $\Pi$, and $w-\Pi u/(1-\Pi)$ with probability $1-\Pi$; $1>\Pi>0$ and $u>0$. To simplify notation, we choose a transformation of the linear function $C(.)$ such that $C(\xi)=\xi$ for all $\xi$. Applying expression (4) and rearranging, we have

$$(5)\quad A_i \succsim A_k \quad \text{iff} \quad Q\left(\frac{\Pi u}{1-\Pi}\right)-\frac{\Pi}{1-\Pi}Q(u)\gtreqless 0.$$

With $Q(\xi)$ convex for all $\xi>0$, it follows that $A_i \succsim A_k$ iff $\Pi \gtreqless 0.5$. Thus, under the special assumption that $C(.)$ is linear, regret theory predicts that individuals will accept small-stake large-prize fair gambles ($\Pi<0.5$) but reject large-stake small-prize fair gambles ($\Pi>0.5$).

TABLE 1

| | $\Pi_P$ | $1-\Pi_P$ |
|---|---|---|
| $P$ | $y_P$ | $x_P$ |
| $P^*$ | $r_P$ | $r_P$ |

TABLE 2

| | $\Pi_\$$ | $1-\Pi_\$$ |
|---|---|---|
| \$ | $y_\$$ | $x_\$$ |
| \$* | $r_\$$ | $r_\$$ |

TABLE 3

| | $\Pi_\$$ | $\Pi_P-\Pi_\$$ | $1-\Pi_P$ |
|---|---|---|---|
| $P$ | $y_P$ | $y_P$ | $x_P$ |
| \$ | $y_\$$ | $x_\$$ | $x_\$$ |

In the light of this result, consider Table 1. If $r_P$ were equal to the expected dollar value of the $P$ bet (which we shall write as $V(P)$) then $P$ (seen in relation to $P^*$) would be a large-stake small-prize fair gamble (recall that $\Pi_P>0.5$). Thus $P \prec P^*$. So to achieve $P \sim P^*$, which must be the case if $r_P$ is a reservation price, it is necessary that $r_P<V(P)$. Now consider Table 2. In this case $\Pi_\$ \leqslant 0.5$, so to achieve $\$ \sim \$^*$ it is necessary that $r_\$ \geqslant V(\$)$. But $V(P)=V(\$)$. Combining these results, $r_P<r_\$$. Thus, under the special assumption of linearity, regret theory explains both the tendency for individuals to place a higher reservation price on the \$ bet, and the tendency for that price to exceed the expected value of the bet.

Now consider the choice between $P$ and \$ represented in Table 3. Applying expression

(4) again, and retaining our assumption about $C(.)$, we have

$$(6)\quad P \gtreqless \$ \text{ iff } \Pi_\$ Q(y_P - y_\$) + (\Pi_P - \Pi_\$) Q(y_P - x_\$) + (1-\Pi_P) Q(x_P - x_\$) \gtreqless 0.$$

Because both bets have the same expected dollar value, we know that

$$(7)\quad \Pi_\$ (y_P - y_\$) + (\Pi_P - \Pi_\$)(y_P - x_\$) + (1-\Pi_P)(x_P - x_\$) = 0.$$

We also know that

(i) $1 > \Pi_\$, \Pi_P - \Pi_\$, 1 - \Pi_P > 0$;
(ii) $y_P - y_\$ < 0$;
(iii) $y_P - x_\$ > 0$;
(iv) $|x_P - x_\$| < |y_P - y_\$|, |y_P - x_\$|$.

Given all this, the assumption that $Q(\xi)$ is convex for all $\xi > 0$ entails the following results (see the Appendix for proof):

Case I: $P \prec \$$ if $x_P - x_\$ > 0$ and $y_\$ - y_P > y_P - x_\$$
Case II: $P \prec \$$ if $x_P - x_\$ < 0$ and $\Pi_P - \Pi_\$ \geqslant 0.5$
Case III: $P \succ \$$ if $x_P - x_\$ < 0$ and $y_\$ - y_P < y_P - x_\$$
Case IV: $P \succ \$$ if $x_P - x_\$ > 0$ and $\Pi_\$ \geqslant 0.5$

These four cases are not exhaustive—there are other cases in which no firm prediction can be made—but they cover all six pairs of gambles in Grether and Plott's experiments; four of those pairs were cases for which $P \prec \$$ is predicted, and two were cases where the prediction is $P \succ \$$.[3] But the most significant feature is that in cases III and IV, preference reversal is predicted, and the form of reversal – $P \succ \$$ and $r_P < r_\$$ – is the one most frequently observed. Of course, if we relax the assumption that $C(.)$ is linear, we can no longer make such clear-cut predictions. Nevertheless, the common form of preference reversal remains consistent with regret theory and with its presumption that individuals are rational, optimizing agents.

## APPENDIX

To simplify notation, let:

$$\Pi_1 \equiv \Pi_\$ \qquad a_1 \equiv y_P - y_\$$$
$$\Pi_2 \equiv \Pi_P - \Pi_\$ \qquad a_2 \equiv y_P - x_\$$$
$$\Pi_3 \equiv 1 - \Pi_P \qquad a_3 \equiv x_P - x_\$.$$

Then the problem is to determine the sign of

$$(\text{A1})\quad \Pi_1 Q(a_1) + \Pi_2 Q(a_2) + \Pi_3 Q(a_3),$$

given that

$$(\text{A2})\quad \Pi_1 a_1 + \Pi_2 a_2 + \Pi_3 a_3 = 0$$

$$(\text{A3})\quad 1 > \Pi_1, \Pi_2, \Pi_3 > 0$$

$$(\text{A4})\quad a_2 > 0 > a_1$$

$$(\text{A5})\quad |a_1|, |a_2| > |a_3|$$

$$(\text{A6})\quad Q(\xi) = -Q(-\xi),$$

and $Q(\xi)$ is convex for all $\xi > 0$.

Case I: $-a_1 > a_2 > a_3 > 0$

In this case, $a_1$ is the only $a_i$ to be negative, and is also the $a_i$ with the largest absolute value. Thus, because of (A2) and (A6), (A1) must be negative, that is, $P \prec \$$.

Case III: $-a_2 < a_1 < a_3 < 0$

In this case, $a_2$ is the only positive $a_i$, and is the $a_i$ with the largest absolute value. Thus (A1) must be positive, that is, $P \succ \$$.

Case II: $a_1, a_3 < 0 < a_2$; $\Pi_2 \geqslant 0.5$

since $Q(\xi)$ is concave for all $\xi < 0$:

$$(\text{A7})\quad \Pi_1 Q(a_1) + \Pi_3 Q(a_3) < (\Pi_1 + \Pi_3) Q\left(\frac{\Pi_1 a_1 + \Pi_3 a_3}{\Pi_1 + \Pi_3}\right).$$

---

[3] See their Table 2. Case I covers pairs of gambles numbers 1, 4 and 5; Case II covers pair number 2; Case III covers pairs numbers 3 and 6.

Consider a choice between (*i*) the certainty of a consequence of zero and (*ii*) the consequences $a_2$, $(\Pi_1 a_1 + \Pi_3 a_3)/(\Pi_1 + \Pi_3)$ with probabilities $\Pi_2$, $\Pi_1 + \Pi_3$. Since (*ii*) is an actuarially fair gamble with the probability of winning, $\Pi_2$, $\geqslant 0.5$, we know from the result concerning fair gambles given in the paper that

$$\text{(A8)}\quad \Pi_2 Q(a_2) + (\Pi_1 + \Pi_3) Q\left(\frac{\Pi_1 a_1 + \Pi_3 a_3}{\Pi_1 + \Pi_3}\right) \leqslant 0.$$

Combining (A7) and (A8):

$$\text{(A9)}\quad \Pi_1 Q(a_1) + \Pi_2 Q(a_2) + \Pi_3 Q(a_3) < 0,$$

that is, $P \prec \$$.

Case IV: $a_2, a_3 > 0 > a_1;\ \Pi_1 \geqslant 0.5$

A similar argument to that given for Case II above establishes that (A1) is positive, that is, $P \succ \$$.

## REFERENCES

**Grether, David and Plott, Charles,** "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

**Lichtenstein, Sarah and Slovic, Paul,** "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, July 1971, *89*, 46–55.

**Lindman, Harold,** "Inconsistent Preferences Among Gambles," *Journal of Experimental Psychology*, August 1971, *89*, 390–97.

**Loomes, Graham and Sugden, Robert,** "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *Economic Journal*, December 1982, *92*, 805–24.

**Schoemaker, Paul,** "The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations," *Journal of Economic Literature*, June 1982, *20*, 529–63.

**Von Neumann, John and Morgenstern, Oskar,** *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1947.

# Aggregate Consequences of Fixed Costs of Price Adjustment

*By* JULIO J. ROTEMBERG*

This paper studies the general equilibrium consequences of monetary growth in an economy in which firms must pay a fixed cost every time they change their price. Hence it presents an extension of the partial equilibrium models of Eytan Sheshinski and Yoram Weiss (1977), and Michael Mussa (1981). A substantial part of this paper is devoted to criticizing Mussa's paper. He claims to have shown that under these conditions the change in the price level depends both on the change in the price level that would have prevailed in the absence of costs of changing prices (the equilibrium rate of change in prices) and on the difference between the actual and the equilibrium price levels. I show that, under his assumptions, the actual and equilibrium price levels always coincide. This is a consequence of his assumption that equilibrium prices grow at a constant rate. Moreover, I argue that in environments like those of Sections I and III of his 1981 paper and those of his 1982 paper in which equilibrium prices do not grow at a constant, his rationale does not justify the use of his formula (18).[1]

Section I of this paper presents the maximization problem of a representative firm, while Section II discusses the ensuing equilibrium.

## I. Firms

The firms in this economy are indexed by $i$ and produce differentiated products. They face identical demand and cost functions. The demand curve facing firm $i$ at time $t$ is

$$(1) \qquad Q_{it} = \left(\frac{P_{it}}{P_t}\right)^{-b}\left(\frac{M_t}{P_t}\right)^{f}; \; i \in [0,1],$$

where $b$ and $f$ are parameters. At time $t$, $Q_{it}$ and $P_{it}$ are, respectively, the quantity demanded and the price of good $i$, while $P_t$ and $M_t$ are the price level and the size of the money stock. The quantity demanded of any good depends not only on its relative price, but also on real money balances. As suggested in my 1982 paper, this dependence will arise if real money balances are a direct source of utility. The price level at $t$ is given by

$$(2) \qquad P_t = exp\left\{\int_0^1 log(P_{it})\, di\right\}.$$

The cost to firm $i$ of producing $Q_{it}$ is

$$(3) \qquad C(Q_{it}) = UP_tQ_{it}^2/2,$$

where $U$ is a parameter. This cost function can arise, as shown in my 1982 paper, both when only goods are required to produce goods and when labor is a factor of production traded in a classical market. Equation (3) guarantees that, in the short run, marginal cost curves slope upwards.

In absence of costs of changing prices, firms would charge a price whose logarithm is $\bar{p}_{it}$:

$$(4) \qquad \bar{p}_{it} = \phi + F(m_t - p_t) + p_t,$$

where lower case letters represent the logarithm of the respective upper case letters. $\phi$ and $F$ are constants given by

$$\phi = \frac{1}{1+b} log\left(\frac{b}{b-1}\right); \; F = \frac{f}{1+b}.$$

As suggested by Mussa (1981) and shown in my earlier paper, the instantaneous loss in profits to firm $i$ from charging a price different from $\overline{P}_{it}$ can be approximated by

---

*Sloan School of Management, Massachusetts Institute of Technology. This paper is a revised version of chapter V of my Ph.D. dissertation presented at Princeton University. I thank my advisor William Branson, as well as Alan Blinder and Robert Porter for helpful comments and suggestions.

[1] It must be noted, however, that Bennett McCallum (1980) presents an alternative rationale for an equation like Mussa's equation (18).

$$B(p_{it}-\bar{p}_{it})^2, \tag{5}$$

where $B$ is equal to $B=\phi b(1+b)\overline{Q}_{it}^2$ and

$$\overline{Q}_{it}=\left(\frac{2bU}{b-1}\right)^{-b/(1+b)}\left(\frac{M_t}{P_t}\right)^{f-bf/(1+b)}.$$

Here, $\overline{Q}_{it}$ is the quantity firm $i$ would sell at $t$ if it charged $\overline{P}_{it}$, and $B$ will be assumed to be constant. In the model studied below, money balances will be constant, therefore $B$ will also be constant. As in Sheshinski and Weiss's paper and Mussa's (1981) paper, firms incur a fixed cost, $A$, every time they change their prices.

Price-setting firms face two types of costs of price adjustment. First, there are the administrative costs which ensue from printing new price lists, informing salesmen of the new prices, etc. Second, there are costs that result from the damage to the reputation of the firm that changes prices often and by large amounts. These latter costs are probably more important in an environment in which the price level moves little, as in the United States.

While the first type of cost is undoubtedly a fixed cost per price change, the second type of cost is, arguably, convex in the size of the price change since customers do not seem to react strongly to small price increases. When the cost of adjustment is convex, a solution to the firm's optimal control problem is guaranteed to exist (see Harold Kushner for a proof) and, furthermore, the solution is easily computed when the cost of adjustment can be approximated by a quadratic function. This approach is taken in my earlier paper.

Instead, only special cases of the control problem when the cost of adjustment is concave can be solved. Here, the objective of the firms is assumed to be the maximization of average real profits per unit of time. This, as shown by Kushner, is equivalent to the limit of the maximization of discounted profits as the discount factor tends to one. Moreover, as in the analysis of Sheshinski and Weiss, and Mussa (1981), I solve the firm's problem only for the case in which the target price, $\bar{p}_{it}$, grows at a constant rate. I will then show that this is a property of the equilibrium when there is a constant rate of monetary growth.

Because there are fixed costs to changing prices, the firms will not change their prices continuously. Instead, they will keep their prices constant for discrete intervals of time. During each interval, they will charge an optimal price that minimizes average losses during the interval. Given the rule for choosing these optimal prices, the intervals themselves can be chosen to minimize losses per unit of time. It is the second part of this optimization that yields computable solutions only in simple cases.

Mussa (1981) considers the case in which $\bar{p}_{it}$ grows at a constant rate, $\overline{\Pi}_i$:

$$\bar{p}_{it}=\bar{p}_{i0}+\overline{\Pi}_i t, \tag{6}$$

where $\bar{p}_{i0}$ is a constant.

As shown by Mussa (1981), firms for which $\bar{p}_{it}$ is given by (6) who minimize average costs per unit of time and keep their price constant between $t$ and $t+T$ will charge the price $p_{it}$ given by

$$p_{it}=\bar{p}_{i0}+\overline{\Pi}_i t+\overline{\Pi}_i T/2 \tag{7}$$

during this interval. An argument similar to Mussa (1981) establishes that, if the firm expects $\overline{\Pi}_i$ to be the rate at which $\bar{p}_{it}$ will grow forever, the value of $T$ which minimizes average losses per unit of time is

$$\overline{T}_i=\left(\frac{12B}{A}\right)^{1/3}\overline{\Pi}_i^{-2/3}. \tag{8}$$

## II. Equilibrium

Following Mussa (1981), I assume that firms are uniformly distributed over the time of their last price change. He also assumes that "all commodities have the same frequency of price change [i.e., the same $T_i$], that price adjustments are evenly spaced and that the rates of change of the equilibrium prices [given by (6)] of all commodities are all the same" (1981, p. 1023). Even though these assumptions imply that $\overline{\Pi}_i$ is the same for all firms, he writes the price level at $t$ as

$$p_t = \frac{1}{T}\int_{t-T}^{t}\left(\bar{p}_i + \frac{T\overline{\Pi}_i}{2}\right)di. \tag{9}$$

He then assumes that $\bar{p}_i + (T/2)\overline{\Pi}_i$ is "close to a linear function of $i$." Hence $\bar{p}_i + (T/2)\overline{\Pi}_i \simeq \alpha + i\overline{\Pi}$, where $\alpha$ and $\overline{\Pi}$ are constants. But, for the rate of equilibrium inflation faced by each firm to be unchanging:

$$\overline{\Pi}_i = \overline{\Pi}_{i+nT};\ \bar{p}_{i+nT} = \bar{p}_i + n\overline{\Pi}_i,$$

where $n$ is any integer. It is easy to establish that these conditions ensure that $\overline{\Pi}_i = \overline{\Pi}$.

It is important to stress that Mussa's assumptions repeatedly imply that $\overline{\Pi}_i = \overline{\Pi}$ because this leads to conclusions which are much stronger than Mussa's. In particular, he shows that

$$dp_t/dt = \overline{\Pi} + \delta(p_t - \bar{p}_t), \tag{10}$$

where $\delta$ is a constant. However, his assumptions imply Proposition 1.

PROPOSITION 1: *When $\overline{\Pi}_i = \overline{\Pi}$ for all $i$ and, without loss of generality $\bar{p}_{i0} = \alpha$, while firms are uniformly distributed over the time of their last price change*:

$$p_t = \bar{p}_t \tag{11}$$

PROOF:

Using (2), (6), and the fact that firms are uniformly distributed over the time of their last price change:

$$p_t = \frac{1}{T}\int_{t-T}^{t} \alpha + \overline{\Pi}u + \frac{T}{2}\overline{\Pi}\Bigg) du$$
$$= \alpha + \overline{\Pi}t = \bar{p}_t.$$

Therefore under Mussa's (1981) assumptions, the price level is equal to the equilibrium price level. This conclusion depends crucially on the constancy of the rate of growth of the equilibrium prices. However, Mussa (1981) provides no evidence that (10) holds for more general paths of $\bar{p}$. Moreover, it is obvious that for a variety of paths of $\bar{p}$, (10) is false. In particular, consider Robert Barro's case in which $\bar{p}$ follows a random walk. Then, when there is a shock to $\bar{p}$, some firms adjust their price because their relative price has now fallen below some critical level $s$. However, after the shock, prices will stay constant until the next shock. There is no continuing tendency for the price level to come back to the equilibrium price level. Moreover, the effect of the shocks on the price level is not simply a function of the average discrepancy $(\bar{p}_t - p_t)$. Instead, it depends on the whole distribution of prices at the time of the shock.

It must be noted that Mussa (1981, 1982) and Bennett McCallum (1978) apply equation (10) to models in which $\bar{p}_t$ does not grow at a constant rate. Mussa's (1981) rationale does not provide any microfoundation for the use of equation (10) in these contexts.

I now establish that (11) is a property of an equilibrium in which $\bar{p}$ is given by (4) and money grows at the rate $\overline{\Pi}$.

PROPOSITION 2. *When the evolution of the money stock is given by $m_t = m_0 + \overline{\Pi}t$ and firms are uniformly distributed over the time of their last price change, an equilibrium price level is given by*

$$p_t = \frac{\phi}{F} + m_0 + \overline{\Pi}t. \tag{12}$$

PROOF:

My 1982 paper establishes that in an economy such as this one, equilibrium in the goods markets is sufficient for overall equilibrium. Here an equilibrium is a perfect foresight path for the price level such that, given that firms expect this path to prevail, their optimal plans are consistent with this equilibrium. At this equilibrium firms are willing to supply the quantities given by (1).

Suppose (12) is an equilibrium. Then, using (4):

$$\bar{p}_{it} = \phi + F\left\{m_0 + \overline{\Pi}t - \left(\frac{\phi}{F} + m_0 + \overline{\Pi}t\right)\right\}$$
$$+ \phi/F + m_0 + \overline{\Pi}t$$
$$= \phi/F + m_0 + \overline{\Pi}t.$$

So the price charged between $i$ and $i+T$ by the firm which changes its price at $i$ is given by (7) with $\bar{p}_{i0}$ equal to $m_0+\phi/F$. Therefore,

$$\begin{aligned} p_t &= \frac{1}{T}\int_{t-T}^{t}\left(\frac{\phi}{F}+m_0+\overline{\Pi}u+\frac{\overline{\Pi}T}{2}\right)du \\ &= \phi/F+m_0+\overline{\Pi}t \end{aligned}$$

Note that, at this equilibrium $(m_t-p_t)$ is constant. This establishes that $B$ is a constant in (5). More importantly, it establishes that the index of aggregate output $q_t$ given by[2]

$$q_t=\int_0^1 q_{it}\,di=f(m_t-p_t)$$

does not depend on the rate of monetary growth $\overline{\Pi}$. It is interesting to note that my 1982 analysis establishes that $q_t$ is also unaffected by the rate of steady monetary growth when firms maximize expected profits per unit of time and face quadratic costs of changing prices.

---

[2] Indices, like $q_t$, which is equal to the logarithm of the product of outputs in all sectors, are widely used (see, for instance, Robert Lucas 1973). They behave like *GNP* when output in the different sectors exhibit a high degree of coherence. Such coherence does indeed seem to characterize business cycles (Lucas, 1977).

## REFERENCES

**Barro, Robert J.,** "A Theory of Monopolistic Price Adjustment," *Review of Economic Studies*, January 1972, *39*, 17–26.

**Kushner, Harold J.,** *Introduction to Stochastic Control*, New York: Holt, Rinehart and Winston, 1971.

**Lucas, Robert E., Jr.,** "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 320–29.

______, "Understanding Business Cycles," in Karl Brunner and Allan H. Meltzer, eds, *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics Suppl.* 1977, 7–27.

**McCallum, Bennett T.,** "A Monetary Ineffectiveness Result in a Model with a Predetermined Price Level," *Economic Letters*, 1978, *3*.

______, "Rational Expectations and Macroeconomic Stabilization: An Overview," *Journal of Money, Credit and Banking*, November 1980, *12*, 716–46.

**Mussa, Michael,** "Sticky Prices and Disequilibrium Adjustment in a Rational Model of the Inflationary Process," *American Economic Review*, December 1981, *71*, 1020–27.

______, "A Model of Exchange Rate Dynamics," *Journal of Political Economy*, February 1982, *90*, 79–104.

**Rotemberg, Julio,** "Monopolistic Price Adjustment and Aggregate Output," *Review of Economic Studies*, October 1982, *49*, 517–31.

**Sheshinski, Eytan and Weiss, Yoram,** "Inflation and Costs of Price Adjustment," *Review of Economic Studies*, June 1977, *44*, 287–304.

# Taxes in a Life Cycle Growth Model with Bequests and Inheritances

By LAURENCE S. SEIDMAN*

In a recent paper, Lawrence Summers (1981) analyzes alternative taxes in a life cycle growth model. This paper extends his analysis to include bequests and inheritances.

This inclusion is worth undertaking for several reasons. First, several recent empirical studies have suggested that bequests may account for an important share of the capital stock (Laurence Kotlikoff and Summers, 1981). Second, it is important to explore whether the comparison of taxes is significantly affected when a bequest motive is introduced. Third, the treatment of bequests under a consumption tax is controversial. The "endowment rationale" calls for treating a bequest as taxable consumption by the deceased (Peter Mieszkowski, 1980); the "real wage rationale" calls for exempting bequests if such exemption raises the steady-state capital intensity and real wage (my 1980 article).

This paper confines itself to positive analysis of steady states. Transition paths, and normative implications, are explored elsewhere (see my 1982 paper).

## I. Bequests, Inheritances, and Taxes in a Life Cycle Growth Model[1]

The representative person is assumed to choose a consumption path from labor force entry (work age zero) to death. All ages are "work ages," expressed relative to a labor force entry age of zero. Upon entry, each person chooses $(C_t, K_J)$ to maximize (1) subject to (2):

$$(1)\qquad U=\int_0^J C_t^{\gamma} e^{-\rho t}\,dt + bK_J^{\gamma} e^{-\rho J}$$

$$\text{if}\qquad \gamma<1,\quad \gamma\neq 0$$

$$U=\int_0^J \ln C_t e^{-\rho t}\,dt + b\ln K_J e^{-\rho J}$$

$$\text{if}\qquad \gamma=0;$$

$$(2)\qquad \int_0^J \frac{C_t}{1-t_c}e^{-r_n t}\,dt + \frac{K_J}{1-t_J}e^{-r_n J} = (1-t_H)H + \int_0^R w_{xt}e^{-r_n t}\,dt,$$

where $r_n=(1-t_r)r$; $w_{xt}=(1-t_w)w_t$; and

$C_t$ = consumption at age $t$,
$J$ = age of death (exogenous),
$K_J$ = bequest at death,
$H$ = inheritance received at work age zero,
$r$ = rate of interest,
$w_t$ = wage at age $t$,
$R$ = age of retirement (exogenous) ($R<J$),
$t_c$ = consumption tax rate (ratio of consumption tax to the sum of consumption plus tax; hence $C_t/(1-t_c)$ is the sum of consumption plus tax),
$t_J$ = bequest tax rate (ratio of bequest tax to the sum of bequest plus tax),
$t_H$ = inheritance tax rate,
$t_w$ = wage tax rate, and
$t_r$ = capital income tax rate.

The parameter $\rho$ is the subjective discount rate; its constancy insures consistency in the sense of Robert Strotz (1956); see also Anthony Atkinson (1971). It can be shown that $\sigma\equiv 1/(1-\gamma)$ ($\sigma>0$) is the elasticity of substitution between $C_t$ and $C_{t'}$, and between $C_t$ and $K_J$, along the optimal path; $\sigma$ will play a key role in the forthcoming analysis. The parameter $b$ is the "taste" for bequests, which is implicitly zero in Summers' paper, but positive here. It can be shown that $b = (K_J/C_J)^{1/\sigma}$; in the $\sigma = 1$ case, simply the ratio of bequest to consumption at the age of death.

---

*Associate professor of economics, University of Delaware. I am grateful to my research assistant, Ken Schwab, who wrote the program for, and executed, the computor calculations.

[1] Derivations for all results presented in this paper are provided in my 1981 working paper, available upon request. For tractability, it is assumed a bequest is given only at death, and an inheritance is received only at labor force entry; the results may be sensitive to this simplification.

It should be emphasized that, in this model, $b > 0$ indicating a "bequest motive," does not necessarily imply that the individual cares about the welfare of his heirs. A person may derive utility from planning a bequest and accumulating the corresponding wealth, because of the security, prestige, or ego gratification that may accompany such wealth accumulation. Alternative modeling of a bequest motive based solely on intergenerational altruism, which focuses on the behavior and utility of heirs, could conceivably yield different results.

The conclusions may prove sensitive to the choice of utility function. On the other hand, (1) is a natural extension of the function utilized by Summers, thereby facilitating comparison; and may be sufficiently flexible to represent any plausible behavior.

As in Summers, it is assumed there is no government saving; the capital stock is accumulated from individual saving. The steady state of the life cycle growth model is obtained as in Summers'; except that the existence of a bequest motive imposes a new relationship. Although each person takes $H$ as given when choosing $(C_t, K_J)$, in the steady state of the growth model, the inheritance received is linked to the bequest chosen:

$$(3) \qquad P_J(v)K_J(v) = P_0(v)\cdot H(v),$$

where $P_t(v)$ = population of age $t$ at time $v$.

Given a population growth rate $n$, and the fact that in the steady state, $w(v)$, $H(v)$, and $K_J(v)$ must all grow at the rate of labor-augmenting technical change $g$, one person's bequest and inheritance are related by

$$(4) \qquad H(v) = e^{-(n+g)J}\cdot K_J(v+J).$$

For a Cobb-Douglas production function, the steady-state $r$ of the economy is given implicitly by

$$(5) \quad r = \left[\alpha(r_n - n - g)/(1-\alpha)(1-t_w)\right] \Big/ \left[\frac{F(g-r_n, R)\left[F(y_n, J) + b^\sigma e^{y_n J}\cdot M\right]}{F(-n, R)D(1-t_c)^{1-\sigma}} - 1\right]$$

where $M \equiv (1-t_c)^{1-\sigma}(1-t_J)^{\sigma-1}(t_J + t_H - t_H t_J)$;

$$D \equiv (1-t_c)^{\sigma-1}F(m_n, J) + (1-t_J)^{\sigma-1} \times b^\sigma e^{m_n J} - (1-t_H)(1-t_J)^\sigma b^\sigma e^{y_n J};$$

$$F(a,c) \equiv \int_0^c e^{at}\,dt \equiv \frac{e^{ac}-1}{a};$$

$$m_n \equiv (r_n - \rho)\sigma - r_n;$$

$$y_n \equiv (r_n - \rho)\sigma - (n+g).$$

A computer-implemented iterative procedure was used to obtain $r$ from (5). Given a Cobb-Douglas function, and neoclassical factor payments according to marginal productivity, then capital per effective labor, $k$, is:

$$(6) \qquad k = [\alpha/r]^{[1/(1-\alpha)]}.$$

## II. The Impact of Taxes

Suppose the tax rate is set under each tax so that steady-state tax revenue per unit of effective labor is equal for all taxes. Six steady states are compared: no-tax, and the following five taxes:

($C$-Tax): A consumption tax that exempts a bequest;
($CB$-Tax): A consumption tax that taxes a bequest;
($W$-Tax): A wage tax that exempts an inheritance;
($WH$-Tax): A wage tax that taxes an inheritance;
($Y$-Tax): An income tax.

A representative sample of results is presented in Table 1 (a more complete set is

TABLE 1—STEADY STATES UNDER ALTERNATIVE TAXES WHEN TAX RATES ARE SET TO ACHIEVE EQUAL TAX REVENUE PER EFFECTIVE LABOR

| $\sigma$ | $\rho$ | | No-Tax | *C*-Tax | *CB*-Tax | *W*-Tax | *WH*-Tax | *Y*-Tax |
|---|---|---|---|---|---|---|---|---|
| 1 | .10 | $r$ | 11.894 | 11.894 | 11.926 | 12.057 | 12.099 | 13.165 |
| | | $s$ | 7.567 | 7.567 | 7.546 | 7.464 | 7.438 | 6.836 |
| | | $K_J/c$ | 1.879 | 2.088 | 1.899 | 1.983 | 2.011 | 1.879 |
| | | $txr$ | 0.000 | 10.000 | 9.841 | 13.281 | 12.990 | 9.654 |
| 2 | .10 | $r$ | 10.847 | 10.817 | 10.868 | 10.926 | 10.953 | 11.997 |
| | | $s$ | 8.297 | 8.320 | 8.281 | 8.237 | 8.217 | 7.502 |
| | | $K_J/c$ | 2.812 | 3.402 | 2.852 | 2.966 | 3.018 | 2.812 |
| | | $txr$ | 0.000 | 10.000 | 9.765 | 13.152 | 12.713 | 9.593 |
| .5 | .10 | $r$ | 13.899 | 13.937 | 13.959 | 14.213 | 14.296 | 15.402 |
| | | $s$ | 6.475 | 6.458 | 6.447 | 6.332 | 6.295 | 5.843 |
| | | $K_J/c$ | 1.512 | 1.604 | 1.528 | 1.593 | 1.615 | 1.513 |
| | | $txr$ | 0.000 | 10.000 | 9.869 | 13.474 | 13.250 | 9.763 |
| .1 | .01 | $r$ | 20.317 | 20.651 | 20.642 | 21.620 | 22.112 | 22.557 |
| | | $s$ | 4.430 | 4.358 | 4.360 | 4.163 | 4.070 | 3.990 |
| | | $K_J/c$ | 1.246 | 1.273 | 1.260 | 1.300 | 1.322 | 1.246 |
| | | $txr$ | 0.000 | 10.000 | 9.885 | 13.932 | 13.842 | 9.932 |

*Notes*: All entries in the table for $r$, $s$, $K_J/c$, and $txr$ are percentages. $t_c$ is set at 10 percent under the *C*-Tax. In each $txr$ row, the tax rates of four other taxes are set so that all five taxes raise the same tax revenue per effective labor in the steady state (error $<.00001$).

The production function is Cobb-Douglas, with the capital share, 30 percent; the labor force growth rate $n$ is 1 percent; the labor-augmenting technical change rate $g$ is 2 percent; the age of entry into the workforce is 0; the age of retirement is 45; the age of death is 60; and the bequest parameter $b$ is 1.6.

$r$ = interest rate (which varies inversely with $k$, the capital-effective labor ratio)
$s$ = saving rate
$K_J/c$ = aggregate bequest to aggregate consumption ratio
$txr$ = tax rate
$\sigma$ = elasticity of substitution in the utility function
$\rho$ = subjective discount rate in the utility function (decimal)

provided in my 1981 paper). Note that from (6), $k$ (not presented in Table 1) varies inversely with $r$.

Three conclusions can be drawn from Table 1. The first is that a bequest motive undermines the "neutrality" of a consumption tax: If $b=0$, any *C*-Tax achieves the same steady state $k$ as no-tax; but if $b>0$, then both a *C*-Tax and a *CB*-Tax in general achieve a $k$ different from no-tax (only a *C*-Tax with $\sigma=1$ maintains the no-tax $k$). This can be verified by inspecting (5). For $b=0$, $t_c$ vanishes.

Now consider $b>0$. For a *C*-Tax ($t_c>0$, all other $t_i=0$), $t_c$ is present on the right-hand side of (5), except in the case where $\sigma=1$. For a *CB*-Tax ($t_c=t_J>0$, all other $t_i=0$), $t_c$ is present on the right-hand side of (5) for all $\sigma$ (including $\sigma=1$). Table 1 illustrates these results. Only for a *C*-Tax with $\sigma=1$ is $r$, and therefore $k$, the same as under no-tax.

Why does $b>0$ undermine neutrality? Summers pointed out that, with $b=0$, in response to the imposition of a *C*-Tax the individual would choose to hold constant $C_t/(1-t_c)$, the sum of consumption plus tax, at each age. The decrease in consumption at each age would match the increase in tax; saving at each age ($S_t$) would remain the same as in the no-tax case. With $b=0$, only consumption yields utility, and it is taxed at the same rate at any age; given the utility function, a proportional reduction in $C_t$ at every age is optimal. It should be noted that this neutrality result depends on the utility function, and on the assumption that the capital stock is accumulated solely from individual saving (government saving is assumed zero).

With $b>0$, it should not be surprising that a $C$-Tax that exempts a bequest generally is not neutral, since both a bequest and consumption now yield utility, but are not taxed symmetrically. It is perhaps more surprising that a $CB$-Tax, which taxes a bequest and consumption symmetrically, also is not neutral, since at first glance this case may seem analogous to the $b=0$ case where neutrality is achieved. Indeed, *given* $H$, it is once again true that the individual would choose to hold $C_t/(1-t_c)$ constant (and $K_J/(1-t_c)$ constant), as can be seen from (7) and (8), obtained by maximizing (1) subject to (2):

$$(7)\qquad \frac{C_t}{(1-t_c)}=\frac{[H+w_0F(g-r,R)]e^{(r-\rho)\sigma t}}{F(m,J)+b^{\sigma}e^{mJ}};$$

$$(8)\qquad \frac{K_J}{(1-t_c)}=\frac{[H+w_0F(g-r,R)]b^{\sigma}e^{(r-\rho)\sigma J}}{F(m,J)+b^{\sigma}e^{mJ}};$$

where $w_0=$ wage at work age zero. (Note that $t_c$ does not appear on the right-hand side of either (7) or (8).)

What, then, undermines neutrality? In (8), $K_J$ would be reduced by the amount of the tax, so the sum stays constant. But through the steady-state link between $K_J$ and $H$, (3), the reduction in $K_J$ would alter $H$, so that $H$ would not remain constant in (7). Thus, while a partial equilibrium analysis ($H$ given) seems to suggest that a $CB$-Tax would be neutral, a general equilibrium analysis shows otherwise.

The second conclusion to be drawn from Table 1 is, as might be conjectured, a consumption tax that exempts bequests ($C$-Tax) usually achieves a higher steady-state $k$ than a consumption tax that taxes bequests ($CB$-Tax), when tax rates are set to equate steady-state revenue per effective labor. Perhaps surprisingly, however, the reverse is possible. Which achieves the higher $k$ depends on the relative strengths of the substitution effect which favors the $C$-Tax, and the postponement effect which favors the $CB$-Tax.

To compare these two taxes, it is useful to compare each to no-tax. Under the $C$-Tax, maximizing (1) subject to (2), and imposing the steady-state link between $H$ and $K_J$, yields

$$(9)\qquad C_t/(1-t_c)=\frac{w_0F(g-r,R)e^{(r-\rho)\sigma t}}{F(m,J)+b^{\sigma}(1-t_c)^{1-\sigma}(e^{mJ}-e^{yJ})};$$

$$(10)\qquad K_J=\frac{(1-t_c)^{(1-\sigma)}w_0F(g-r,R)b^{\sigma}e^{(r-\rho)\sigma J}}{F(m,J)+b^{\sigma}(1-t_c)^{(1-\sigma)}(e^{mJ}-e^{yJ})}.$$

With $b>0$, (9) and (10) show that $C_t/(1-t_c)$ and $K_J$ would be independent of $t_c$ only in the special case $\sigma=1$. Computer calculations of $r$ in (5) show that if $\sigma<1$, a $C$-Tax would achieve a lower $k$ (higher $r$) than no-tax (see Table 1); but if $\sigma>1$, a $C$-Tax would actually achieve a higher $k$ than no-tax. Why is this so? A $C$-Tax that exempts a bequest induces a shift out of consumption into a bequest. When $\sigma=1$, the magnitude of the shift is such that $C_t/(1-t_c)$ and $K_J$ (and thus $H$) are unaffected by the imposition of the tax ($C_t$ falls relative to $K_J$), so $S_t$ is unaffected. But if $\sigma>1$, a greater shift occurs, so that $S_t$ actually increases, raising steady-state $k$.

Under the $CB$-Tax, maximizing (1) subject to (2), and imposing the steady-state link between $H$ and $K_J$, yields

$$(11)\qquad C_t/(1-t_c)=\frac{w_0F(g-r,R)e^{(r-\rho)\sigma t}}{F(m,J)+b^{\sigma}\left[e^{mJ}-(1-t_c)e^{yJ}\right]};$$

$$(12)\qquad K_J/(1-t_c)=\frac{w_0F(g-r,R)b^{\sigma}e^{(r-\rho)\sigma J}}{F(m,J)+b^{\sigma}\left[e^{mJ}-(1-t_c)e^{yJ}\right]}.$$

For all $\sigma$, including $\sigma=1$, $C_t/(1-t_c)$ varies with $t_c$. In every case computed, a $CB$-Tax reduced $k$ relative to no-tax. Hence, for $\sigma \geqslant 1$, the $C$-Tax achieved a higher $k$ than the $CB$-Tax (see Table 1).

Even for $\sigma < 1$, for parameter values yielding a plausible $r$, the $C$-Tax achieved the higher $k$. But it should also be noted that for low values of $\sigma$ that yield implausibly high values of $r$ (given a nonnegative $\rho$), the $CB$-Tax achieved the higher $k$. Although it is therefore likely that, as might be conjectured, exempting bequests would raise $k$, it is instructive to understand why this need not be the case for sufficiently small $\sigma$.

The intuition that exempting bequests should raise $k$ derives from the substitution effect—the exemption should induce a shifting out of consumption into bequest. Consider, however, the limiting case in which $\sigma \rightarrow 0$, where no shifting would occur. To raise the same revenue, the $C$-Tax would require a higher tax rate than the $CB$-Tax because its base is smaller due to the exclusion of bequests. Under the $CB$-Tax, a person would pay less tax before death than under the $C$-Tax, which must collect all tax while the person lives and consumes. The postponement of tax under the $CB$-Tax would in itself tend to enable the person to achieve a higher wealth accumulation path than under the $C$-Tax. As $\sigma$ increases from 0, at some point before reaching 1, the substitution effect would outweigh the postponement effect, and the $C$-Tax would achieve the higher $k$.

The third conclusion to be drawn from Table 1 is as follows: when tax rates are set to equate steady-state revenue per unit of effective labor, for all parameters tried yielding a plausible steady-state interest rate and bequest-consumption ratio, the ranking of taxes, beginning with the tax achieving the highest capital intensity, is: 1) $C$-Tax; 2) $CB$-Tax; 3) $W$-Tax; 4) a wage tax that taxes inheritance ($WH$-Tax); 5) an income tax ($Y$-Tax). This ranking results from the interaction of postponement, substitution, and incentive effects.

As Summers pointed out, the postponement effect is the reason that a consumption tax achieves a higher steady-state $k$ than a wage tax when tax rates are set to equate steady-state revenue per effective labor. This effect continues to apply when $b > 0$. Hence, a $CB$-Tax is not equivalent to a $WH$-Tax; the $CB$-Tax achieves the higher $k$ (see Table 1). Also, the postponement effect explains why the $W$-Tax achieved a higher $k$ than the $WH$-Tax.

Finally, the $Y$-Tax postpones tax over the life cycle more than the $W$- or $WH$-Tax. Yet for all parameters tried, the $Y$-Tax achieved the lowest $k$ (as Summers found for $b = 0$). The reason is that, under the $Y$-Tax, the incentive effect of a lower $r_n$ on saving evidently outweighed the postponement effect for all parameters tried. The incentive and postponement effects both combine to reduce the $Y$-Tax $k$ below the $C$-Tax $k$.

## REFERENCES

**Atkinson, Anthony B.**, "Capital Taxes, the Redistribution of Wealth, and Individual Saving," *Review of Economic Studies*, April 1971, *38*, 209–28.

**Kotlikoff, Laurence and Summers, Lawrence,** "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

**Mieszkowski, Peter,** "The Advisability and Feasibility of an Expenditure Tax System," in H. Aaron and M. Boskin, eds., *The Economics of Taxation*, Washington: The Brookings Institution, 1980, 179–201.

**Seidman, Laurence,** "The Personal Consumption Tax and Social Welfare," *Challenge*, September 1980, *23*, 10–16.

______, "Taxes in a Life Cycle Growth Model with Bequests and Inheritances," Working Paper, 1981.

______, "Conversion to a Consumption Tax: The Transition in a Life Cycle Growth Model," Working Paper, 1982.

**Strotz, Robert,** "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, No. 3, 1956, *23*, 165–80.

**Summers, Lawrence,** "Capital Taxation and Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, *71*, 533–44.

# Collateral in Credit Rationing in Markets with Imperfect Information: Note

*By* HILDEGARD C. WETTE*

In their 1981 article, Joseph Stiglitz and Andrew Weiss analyze adverse selection and incentive effects in the loan market. The models considered are based on two crucial assumptions: borrowers are subject to limited liability; and lenders cannot distinguish borrowers (projects) of different risk. Stiglitz and Weiss show that a bank that raises its interest rate may suffer adverse selection because only risky borrowers will be willing to borrow at the higher rate. Thus lenders may choose not to raise the interest rate to eliminate excess demand, resulting in the possibility of a "credit rationing equilibrium." Stiglitz and Weiss also consider briefly the role of collateral in such credit rationing models. They conclude that lenders may choose not to use collateral requirements as a rationing device. An increase in collateral requirements, like an increase in the interest rate, potentially leads to a decrease in the lender's expected return on loans because of resulting adverse incentive and selection effects.

The purpose of this note is to further investigate the role of collateral in these models. Stiglitz and Weiss' discussion in Section III establishes that adverse selection effects can result from increases in collateral when borrowers are risk averse. I will show by returning to a model they discussed earlier in Section I, that the adverse selection effects can also occur when borrowers are risk neutral.

Stiglitz and Weiss outline a model to consider the use of collateral as a rationing device (Section III). In that model, all potential borrowers face the same array of risky projects; each potential borrower chooses (at most) one of those projects to undertake. The individuals are, by assumption, risk averse with decreasing absolute risk aversion, and possess different amounts of initial wealth. Thus, choice of project (if any) to undertake and the method of finance—self-finance from initial wealth or loan finance—will differ from one individual to the next. Stiglitz and Weiss show that, among those who undertake risky projects and who choose borrowing as the method of finance, wealthier individuals undertake riskier projects. An increase in collateral has two effects on the market for loans: those individuals who remain in the market will choose to undertake less-risky projects; and those individuals who drop out of the market are less-wealthy, low-risk borrowers. If the second effect is sufficiently strong, then increased collateral requirements will mean decreased expected returns for the lender. Thus, a credit rationing equilibrium may occur, since lenders may not choose to use collateral requirements (or the interest rate) to eliminate excess demand. The adverse selection effect just described does not occur in this model if the individuals are risk neutral. Consequently, the potential for a credit rationing equilibrium is limited to cases where borrowers are risk averse.

To see that increases in collateral requirements can also result in adverse selection if borrowers are risk neutral, consider the model Stiglitz and Weiss used to analyze the adverse selection effects of increases in the interest rate (Section I, pp. 395–99). It differs from the Section III model discussed above in three ways. First, borrowers are risk neutral; second, all projects ar loan financed. The analogy to the Section III model is that no individuals have sufficient wealth to self-finance projects.[1] Third, in the Section

---

*J. L. Kellogg Graduate School of Management, Northwestern University and University of Illinois at Chicago. I thank Paul R. Milgrom for helpful comments.

[1] As in all the Stiglitz-Weiss models, the choice of self-financing a portion of the project and loan financing the remainder is ruled out. It will become evident in what follows, however, that individuals are assumed to have enough wealth for the collateral required on loans.

III model, all borrowers face the same array of potential projects, and each borrower is "allowed" to undertake only one of them. In the Section I model, borrowers do not necessarily face the same set of project options, nor do they necessarily undertake only one. The lender knows that all available projects that yield nonnegative expective profit to borrowers at the given interest rate and collateral will be undertaken. (However, as in the Section III model, the lender cannot distinguish *which* borrower is undertaking *which* project.)

Each available project has a distribution of gross payoffs $F(R,\theta)$, where greater $\theta$ denotes greater risk. The lender is assumed to be able to evaluate the mean payoff of any project, but not its riskiness. The distribution of the riskiness of available projects with a given mean is denoted $G(\theta)$.[2] Stiglitz and Weiss show that, because of the limited liability assumption, the riskiest projects are the most profitable from the borrower's point of view. As the interest rate increases, the less-risky projects become unprofitable so borrowers do not undertake them. This result ultimately leads to the possibility of a credit rationing equilibrium.

In what follows, I use this model to show that increases in collateral requirements lead to adverse selection effects. Analysis parallels the Stiglitz-Weiss analysis concerning adverse selection effects of interest rate increases. My theorems differ from their theorems because collateral $C$, rather than the interest rate $r$, is allowed to vary.[3]

LEMMA: *Expected profit to the borrower is an increasing function of the riskiness of the project undertaken.*

PROOF:

As in Stiglitz and Weiss, the limited liability assumption implies that borrower profits on a particular project are a convex function of the gross payoff $(R)$ of the project. Thus, increases in risk $(\theta)$ increase the expected profits of the borrower.

THEOREM 1-m: *For given collateral $C$, there is a critical value $\tilde{\theta}$ such that an individual borrows from the bank (and undertakes the project) if and only if $\theta \geqslant \tilde{\theta}$.*

PROOF:

This follows directly from the lemma. Since the borrower's expected profits are an increasing function of $\theta$, there is a level of $\theta$ $(\tilde{\theta})$ that yields zero expected profit. All projects with $\theta \geqslant \tilde{\theta}$ will thus be undertaken.

THEOREM 2-m: *As the collateral requirement increases, the critical value of $\theta$ below which individuals do not apply for loans increases.*

PROOF:

The value $\tilde{\theta}$ for which the borrower's expected profit is zero satisfies

$$\text{(5-m)}\quad \pi(\tilde{C},\tilde{\theta}) = -\tilde{C}\cdot F(1+r-\tilde{C},\tilde{\theta}) + \int_{1+r-C}^{\infty} [R-(1+r)]\, dF(R,\tilde{\theta}) = 0.$$

Total differentiation yields

(6-m)

$$\frac{d\tilde{\theta}}{d\tilde{C}} = -\left(\frac{\partial\pi(\tilde{C},\tilde{\theta})}{\partial C}\right)\Big/\left(\frac{\partial\pi(\tilde{C},\tilde{\theta})}{\partial\tilde{\theta}}\right)$$

$$= F(1+r-\tilde{C},\tilde{\theta})\Big/\left(\frac{\partial\pi(\tilde{C},\tilde{\theta})}{\partial\tilde{\theta}}\right) > 0.$$

(We know from the lemma that $\partial\pi(\tilde{C},\tilde{\theta})/\partial\tilde{\theta} > 0$; also $F(1+r-\tilde{C},\tilde{\theta}) > 0$. Thus $d\tilde{\theta}/d\tilde{C} > 0$.)

Theorem 2-m is the crucial result. Increases in collateral will result in the less-risky (low $\theta$) borrowers (projects) dropping out of the market for loans. The intuition is as follows. For a given project (given $\theta$), an increase in collateral unambiguously increases the cost—and decreases the profit—to the borrower for some realizations of $R$. (Figure 1 illustrates this result.) Conse-

---

[2] More precisely, greater $\theta$ corresponds to greater risk in the sense of mean preserving spread as defined by Michael Rothschild and Stiglitz.

[3] The notation "m" is used to denote numbers of theorems and equations in Stiglitz-Weiss that have been altered due to my modifications. Clearly, in this analysis it must be assumed that all values of $C$ under consideration are less than $(1+r)$; otherwise the lender would assume no risk in financing projects.

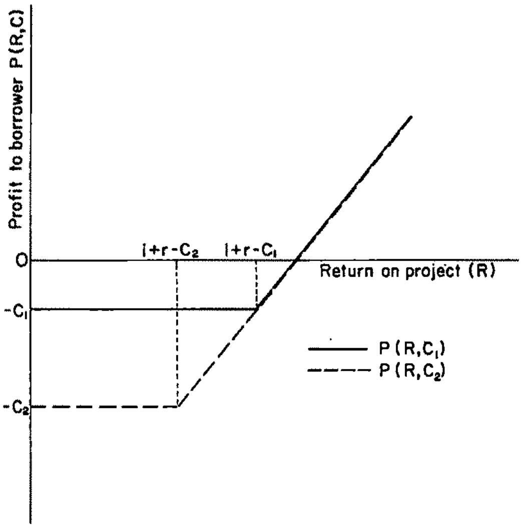


FIGURE 1

quently, expected profit for that project declines. This decrease in expected profit on each project means that some projects that were profitable at the initial level of $C$ become unprofitable at the new, higher level of $C$. From the discussion of Theorem 1-m, we know those are the low-risk projects.

It remains only to show that this adverse selection effect from increases in collateral may cause a decrease in the lender's expected return on loans. The lender's expected return on a loan (project) of risk $\theta$ is $\rho(\theta,\tilde{C})$. Since the lender cannot determine which borrower undertakes the project $\theta$, she must calculate an "average" expected return, $\bar{\rho}(\tilde{\theta}(\tilde{C}),\tilde{C})$.

(4b-m)

$$\rho(\theta,\tilde{C})=\int_0^{1+r-\tilde{C}}(R+\tilde{C})\,dF(R,\theta)$$

$$+(1+r)\cdot(1-F(1+r-\tilde{C},\theta));$$

$$\text{(7-m)}\quad \bar{\rho}(\tilde{\theta}(\tilde{C}),\tilde{C})=\frac{\int_{\tilde{\theta}}^{\infty}\rho(\theta,\tilde{C})\,dG(\theta)}{1-G(\tilde{\theta})}.$$

We want to show that $\bar{\rho}$ is not necessarily an increasing function of $C$. This follows from differentiation of (7-m):

$$\text{(8-m)}\quad \frac{d\bar{\rho}}{d\tilde{C}}=\frac{\int_{\tilde{\theta}}^{\infty}F(1+r-\tilde{C},\theta)\,dG(\theta)}{1-G(\tilde{\theta})}$$

$$+\frac{g(\tilde{\theta})}{1-G(\tilde{\theta})}(\bar{\rho}-\tilde{\rho})\frac{d\tilde{\theta}}{d\tilde{C}},$$

where $\tilde{\rho}=\rho(\tilde{\theta}(\tilde{C}),\tilde{C})$ is the expected return on the least-risky project undertaken. The first term in this expression is positive. The second term is negative because $d\tilde{\theta}/d\tilde{C}>0$ (by Theorem 2-m) and $\bar{\rho}<\tilde{\rho}$. The latter follows from Theorem 3 in Stiglitz-Weiss—the expected return on a loan to the lender is a decreasing function of the riskiness of the project financed by that loan; thus, the expected return to the lender on the least risky loan $\tilde{\rho}$ is greater than the expected return on the "average" loan $\bar{\rho}$ and $d\tilde{\theta}/d\tilde{C}>0$ (by Theorem 2-m).

Since $\bar{\rho}$, $\tilde{\rho}$, and $\tilde{\theta}$ do not depend on $g(\tilde{\theta})$, the negative term can be made arbitrarily large in absolute value by choosing $g(\tilde{\theta})$ large. Then, an increase in collateral reduces the lender's average expected return from loans.[4]

---

[4] The expressions for $d\tilde{\theta}/d\tilde{C}$ (equation (6-m)) and for $d\bar{\rho}/d\tilde{C}$ (equation (8-m)) look very similar to the analagous equations (6) and (8) in Stiglitz-Weiss for changes in $r$:

$$\text{(a)}\quad \frac{d\hat{\theta}}{d\hat{r}}=\frac{1-F((1+r)-C,\hat{\theta})}{\partial\pi(\hat{r},\hat{\theta})/\partial\hat{\theta}};$$

$$\text{(b)}\quad \frac{d\bar{\rho}}{d\hat{r}}=\int_{\hat{\theta}}^{\infty}\frac{[1-F(1+\hat{r}-C,\theta)]\,dG(\theta)}{1-G(\hat{\theta})}$$

$$+\frac{g(\hat{\theta})}{1-G(\hat{\theta})}(\bar{\rho}-\hat{\rho})\frac{d\hat{\theta}}{d\hat{r}}.$$

The similarities occur because changes in $C$ and in $r$ both affect the cost of a given project to the borrower for certain realizations of $R$. In approximate terms, an increase in $r$ affects the cost when the project "succeeds" (hence terms involving the upper tail of the project's distribution of returns, $1-F((1+r)-C,\theta)$); an increase in $C$ affects the cost when the project "fails," hence terms involving the lower tail of the distribution, $F((1+r)-C,\theta)$.

The conclusion in the Stiglitz-Weiss paper is that if borrowers are risk neutral and subject to limited liability, then a lender may not be willing to increase the interest rate on loans to eliminate excess demand; if the borrowers are risk averse, he may also not be willing to raise collateral requirements to eliminate excess demand. The analysis here demonstrates that the lender also may not be willing to use collateral requirements as a rationing device even when borrowers are risk neutral, because increases in collateral can lead to adverse selection effects that decrease the lender's expected return on loans.

## REFERENCES

**Rothschild, Michael and Stiglitz, Joseph E.**, "Increasing Risk: I, A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

**Stiglitz, Joseph E. and Weiss, Andrew**, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–409.

# Labor Supply and Tax Rates: A Correction of the Record

*By* JAMES GWARTNEY AND RICHARD STROUP*

When analyzing the theoretical implications of a change in tax rates, economists often apply the work-leisure analysis for the individual to the aggregate labor market. In this note, we will view the impact of changes in tax rates in a general equilibrium model. We will show that generalizations drawn from the work-leisure analysis for the individual about the impact of a tax cut on labor supply are typically invalid. In addition, the generalizations conceal the true nature of the income effect emanating from a change in tax rates. Section I briefly describes the traditional analysis and indicates the conditions under which it can be appropriately applied to the aggregate labor market. Section II describes the widespread view of the linkage between tax rates and the aggregate supply of labor, which we believe to be incorrect. Section III considers the implications of the changes in tax rates in a general equilibrium framework. Section IV applies this framework to transfer payments in order to better illustrate the fallacy in the work of many writers.

## I. Appropriate Application of the Individual Analysis to the Aggregate Labor Market

Economists have long recognized that a higher real wage rate generates two conflicting influences on the labor supply decision of a worker. First, a higher real wage increases the opportunity cost of leisure, inducing the worker to substitute away from leisure and toward work.[1,2] Other things constant, this substitution effect increases the quantity of labor supplied as real wages rise. Second, the increase in real wages expands the "consumption possibilities" set available to a worker. This income effect permits a worker to increase his consumption of all goods, including leisure. If leisure is a normal good, the income and substitution effects exert opposite impacts on hours worked. Thus, the linkage between real wage rates and amount of labor supplied is ambiguous.

In aggregate, the consumption possibilities of an economy cannot expand unless the production possibilities have expanded. Investments in human capital, technological advancements, and physical capital formation have historically led to growth in output per man-hour and real wage rates. In this instance, it is proper to reason from the individual to the aggregate. Economic growth makes it possible for individuals, *in aggregate*, to attain a higher income, even though they do not increase their hours worked. With economic growth, the combination of more leisure and more income is perfectly consistent with the production possibilities of the economy. It is also proper to reason from the individual to the aggregate *for a specific group* if subsidy programs increase both the pay rates and the real income of the group. The cross-sectional empirical analysis

---

*Florida State University and Montana State University, respectively. We express our appreciation to Thomas McCaleb, Paul Downing, and an anonymous referee for their helpful suggestions. This research was supported in part by The Center for Political Economy and Natural Resources, Montana State University.

[1] For simplicity of exposition, we follow the traditional treatment of leisure simply as hours not worked. A more thorough treatment would allow other dimensions of job choice to enter also. In addition to changing their hours worked as pay rates increase, workers may also be more willing to change locations, work more intensively, accept more responsibility, reduce absenteeism, or pursue other actions that increase their productivity per hour worked. These substitutions of money for nonmoney job characteristics may each change differently when the net wage rate changes. For example, higher take-home pay associated with a potential promotion might lead to more productivity *despite* fewer hours worked, as a worker seeks greater money income *and* more time off to enjoy it.

[2] The higher real wages will also increase the probability that nonworkers will enter the work force. Thus, the hours worked of some individuals will rise from zero to a positive amount.

of experimental income maintenance programs is illustrative of this procedure.[3]

## II. Tax Rates and Labor Supply— A Widespread View

When analyzing the impact of a change in tax rates, many economists have applied the income-leisure analysis of the individual to the aggregate labor supply curve. According to this view, a reduction tax in rates will increase the real income of individuals by the amount of the tax cut. This increase in real income will induce individuals to consume more leisure and less work, at least partially, if not entirely, offsetting the substitution effect. Thus, even the directional impact of a reduction in tax rates on the quantity of labor supplied is indeterminant. The analysis is symmetrical. An increase in tax rates will reduce real income by the amount of the tax increase. Again, the income and substitution effects exert a conflicting influence on hours worked.

Should the reader feel that we have incorrectly outlined the view of those who apply the individual analysis to the aggregate labor market when analyzing the impact of a change in tax rates, we feel compelled to offer a few typical examples of our reasoning. The words of Lyle Gramely, a governor of the Federal Reserve System, nicely illustrates the widespread view. While discussing various proposals to reduce tax rates, Gramley stated:

> It is difficult to predict whether a completely rational economic man would work more or less if taxes were lowered. Lower rates of taxation increase the take-home pay that can be earned from an additional hour of work or a second job, but they also make it possible to attain any given standard of living with less work. [p. 7]

Policymakers are not the only ones who apply the individual analysis to the aggregate economy. In a review article on supply-side economics, James Barth (1981) recently stated:

> The predicted impact of marginal tax reductions on work effort, however, is theoretically questionable. As a result of an increase in after-tax real wage rates, people may work less, not more. The reason is that with higher after-tax real wages, an individual may be able to work fewer hours, while still maintaining or even improving upon his real income or standard of living. This means that although the tax cut will change the relative price of work vis-à-vis leisure and thus induce a "substitution effect" of work for leisure, this may be more than offset by the "income effect" that induces people to work less and devote more time to leisure. [p. 5]

In his analysis of the expected effects of a tax cut on labor supply, Jerry Hausman (1981) stated:

> Income taxes have two effects on labor supply. Taxes lower the net wage and reduce labor supply by the compensated substitution effect. But taxes also have an income effect, which causes individuals to work more since they have been made worse off by the tax. The two effects have opposite signs and might well approximately cancel, causing only a small net effect on labor supply from income taxation. [p. 175]

## III. Tax Rates and Labor Supply— A Reconstructed View

The traditional view is entirely correct for an individual (or group of individuals) whose real income increases by the amount of the tax cut. However, when analyzing the impact of a tax cut, the generalization of the individual analysis to the economy as a whole is invalid. A tax cut does not increase aggregate real income by the amount of the tax reduction.[4] It does not per se change the technical production possibilities of the economy.[5] It will be impossible for *all* individuals to consume both more goods and more leisure as the individual work-leisure analysis implies. Since a reduction in tax rates does not increase aggregate real income by the amount of the tax cut, it is fundamentally different from the case of an increase in real wage rates (and real income) over time and the partial equilibrium income maintenance situation mentioned in Section I.

---

[3] See, for example, the discussion of experimental and nonexperimental data analyses in Michael Keeley (1981), chs. 4 and 5.

[4] If this were true, as those who generalize from the individual work-leisure analysis imply, the real income of all worker's could be increased by reducing personal income tax rates. Clearly, the optimal income tax rate would be zero.

The traditional analysis is deficient in that it fails to consider how changes in tax receipts will influence the availability of public sector goods and services. A simple private-public sector general equilibrium model can be used to clarify this point. Consider a mixed economy of homogeneous individuals. Private goods are produced in the market sector and financed by after-tax income. Public goods are supplied by the government and financed from tax receipts. Individuals derive utility from their consumption of both private and public goods. The resources of the economy are fully employed and, initially, we assume that the government is supplying the optimal quantity of public goods.

In approaching this topic, it is important to carefully specify the *ceteris paribus* conditions and to consider how tax changes alter the revenues available for public sector expenditures. In order to simplify the analysis, we will assume that (a) the monetary authorities hold the supply of money constant, and (b) the size of the budget deficit or surplus is unchanged.[6] These two assumptions will permit us to focus on the "pure tax-cut effect," rather than bringing changes in monetary and fiscal policy into the analysis.

Independent of any additional output emanating from the substitution effect, the reduction in tax rates will clearly lead to a decline in tax revenues. For our general equilibrium model, any expansion in expenditures on private goods stemming from the alleged income effect of the tax cut will be exactly offset by a decline in expenditures on public goods (because of the decline in tax revenues which finance them). When the initial level of government expenditure is socially efficient, then at the margin, the additional utility derived from an expansion in consumption of private goods is exactly offset by the decline in utility associated with the reduction in consumption of public goods. The alleged income effect of the tax cut simply disappears. Only the substitution effect remains and the tax cut will unambiguously increase the quantity supplied of labor. For a tax cut of discrete value, an individual's valuation of the public goods foregone as the result of the income effect of the tax cut will exceed his or her valuation of the additional private goods now available. Real income will fall because government spending is now below the optimal level. Under these circumstances, the negative income effect will reinforce the substitution effect. Unambiguously, the quantity supplied of labor will increase. The economy's production possibilities (for private and public goods) will expand due to the increase in the supply of labor, although the "aggregate utility" of the community will have declined.[7]

---

[5] The fallacy committed by those who apply the work-leisure analysis for the individual to the aggregate labor market when analyzing the expected impact of a tax cut is very similar to that of earlier economists who applied the construction of a money income constant individual demand curve to the aggregate economy. See Milton Friedman (1949) and Martin Bailey (1954) for discussion of this issue. Also see Friedman (1976) for an illustration of a similar fallacy concerning the welfare effects of an excise tax.

[6] Alternatively, one might choose to hold the level of government expenditures constant. If this were the case, the government would have to either print money or borrow in order to obtain the resources necessary to maintain the pretax level of public goods. Given our assumption of full employment, neither of these options would change the production possibilities of the economy. Both would lead to the same conclusions, with regard to the income effect, as our present model.

[7] The increase in aggregate output stemming from the substitution effect may moderate the decline in tax revenue associated with the reduction of rates. Thus, when the *combined* impact of both the substitution and income effects is considered, the decline in expenditures on public goods may be less than the increase in expenditures on private goods. However, this outcome is *contingent* upon the substitution effect inducing an expansion in the supply of labor. Clearly, it would be impossible for this expansion in output (and measured real income) to induce individuals to supply less labor at a lower tax rate as the individual work-leisure analysis implies.

If we relax our assumption that the initial output of public goods was optimal, the income effect associated with a discrete tax cut might be either positive or negative within our general equilibrium model. We have already pointed out that the income effect will be negative if individuals value the public goods that must be foregone as the result of the tax cut more than the expansion in private goods stemming from their increase in disposable income. In this instance, the income and substitution effects reinforce each other. Alternatively, if the initial output of public goods is larger than the optimal, the utility derived from the expansion in consumption of private goods will exceed the utility foregone by the decline in output of public goods as a result of the tax cut. In this instance, a tax cut will increase real income, leading to income and substitution effects that are in opposite directions.

In a general equilibrium setting, the income effect is determined by the individual's valuation of the expansion in private goods relative to the decline in public goods, independent of changes in output emanating from the substitution effect. It reflects the improvement or deterioration in allocative efficiency as resources are shifted between the private and public sectors. In general, the income effect is not equal to the individual's increase in disposable income derived from a tax reduction. The only case where the aggregate income effect is equal to the increase in disposable income, as the individual work-leisure analysis of a tax cut implies, is when individuals place zero value on the public goods that must now be foregone as the result of the shift in output from the public to the private sector.

The general equilibrium approach can also be utilized to show that the traditional view of a tax increase is invalid. Generalizing from the individual work-leisure analysis, the traditional view emphasizes that an increase in tax rates will reduce real income by the amount of the tax increase. According to this view, individuals may choose to work more (even though the opportunity cost of leisure has fallen) in order to make up for the loss of real income. Our general equilibrium model indicates that this view of the income effect is erroneous. Independent of output changes flowing from the substitution effect, if individuals value the expansion in output of public goods emanating from the tax increase more than they value the private goods output that must now be foregone, a positive income effect will reinforce the substitution effect. Unambiguously, the tax increase will cause an increase in consumption of leisure and a decline in work effort.

Alternatively, if the additional output of public goods is valued less than the decline in output of private goods, the income effect will be negative. Under these circumstances, the income and substitution effects will exert a conflicting influence on work effort. However, even in this case, the negative income effect will not be equal to the decline in the individual's disposable income, as the traditional work-leisure analysis implies.

## IV. Tax Rates, Tax Revenues, and Transfer Payments

Perhaps an increase in transfer payments financed by higher tax rates provides the clearest illustration of why it is important to view the income and substitution effects in a general equilibrium setting. This combination—higher taxes and larger transfer payments—is of growing significance. As Table 1 illustrates, both income taxes and transfer payments have expanded as a proportion of national income in recent years. In 1965, income and employment taxes took 15 percent of national income. By 1979, the comparable figure had risen to 23.4 precent. Transfer payments also rose sharply during the period, expanding from 6.6 percent of national income in 1965 to 14 percent in

TABLE 1—THE GROWTH OF INCOME, TAX REVENUE, AND TRANSFER PAYMENTS, 1965–79[a]

| Year | Income and Employment Tax Revenues | Government Transfer Payments |
|---|---|---|
| 1965 | 15.0 | 6.6 |
| 1970 | 19.6 | 9.5 |
| 1975 | 20.7 | 14.0 |
| 1979 | 23.4 | 12.6 |

*Source: Economic Report of the President*, Tables B-17, B-73, and B-74; 1980, and Tax Foundation, Table 11.

[a] Shown as a percentage of national income.

1975, before falling back to 12.6 percent in 1979.

If the government raises tax rates, using the additional tax revenues to finance an equal expansion in transfer payments, the change in aggregate disposable income will be zero. The reduction in the disposable income of each taxpayer is just offset by an equal expansion in transfer payments which increases the disposable income of other individuals.[8] Unless taxpayers as a group have a different income elasticity of demand for leisure than the recipients of the income transfers, the income effect of the tax-transfer program will leave both the aggregate supply of labor and the consumption of leisure unchanged. Of course, the substitution effect will still be present. Unambiguously, it will induce individuals to work less, although how much less is strictly an empirical issue.

If the additional transfer payments are inversely linked to income, they will also increase the implicit marginal tax rate of recipients. For example, this will be true if larger benefits are supplied to lower income recipients while benefit levels decline as the recipient's income approaches the break-even or zero transfer level. Under these circumstances, there is a dual disincentive effect on labor supply. While the income transfers do not *directly* reduce total income, the substitution effect associated with the transfer will induce *both* the taxpayer-donors and the transfer recipients to reduce their work effort.

## V. Summary and Conclusions

Our analysis indicates that it is invalid to generalize from the individual work-leisure analysis to the economy as a whole when analyzing the impact of changes in tax rates on the supply of labor. The change in an individual's disposable income due to a change in taxes is *not* indicative of the change in real income *in aggregate* unless, at the margin, government expenditures produce nothing of value. Independent of any additional output emanating from the substitution effect, an increase in the disposable income of individuals due to a reduction in tax rates will lead to an equal reduction in tax revenues (and expenditures on public sector goods). If, at the margin, a dollar of additional government expenditure is valued as much as a dollar of expenditure in the private sector, transfer of income between the two sectors will leave the real income of citizens unchanged. Under these circumstances, there is no reason to believe that there is a positive income effect which will reduce work effort and tend to offset the positive impact of the substitution effect on the supply of labor. Therefore, the tax cut will increase the quantity supplied of labor.

The analysis is symmetrical. Real income will not decline in response to a tax increase, unless the expansion in government expenditures is valued less than the loss of income from private goods that must now be foregone. Analysis of the income transfer issue illustrates the major point of this paper. If tax rates are increased in order to enlarge transfer payments, *in aggregate*, there will be no income effect. Only the substitution effect remains and it will decrease the quantity supplied of labor under the assumptions of our general equilibrium model.

Of course, the sensitivity of labor supply to a change in tax rates is still an empirical question. The supply of labor may be relatively unresponsive to changes in the price of leisure relative to work. However, the relevant issue is the sensitivity of labor supply to the relative price change, rather than the presence of an alleged "income effect" which reflects changes in after-tax income while ignoring changes in tax revenues and public sector expenditures.

## REFERENCES

**Bailey, Martin J.** "The Marshallian Demand Curve," *Journal of Political Economy*, June 1954, *62*, 255–61.

**Barth, James R.,** "The Reagan Program for Economic Recovery: Economic Rational," *Economic Review*, Federal Reserve Bank

---

[8] Here, for the sake of simplicity, we assume that the transaction cost of the income transfers is zero. We also ignore a potential change in utility that taxpayers may derive from the knowledge that their tax dollars went to a transfer recipient.

of Atlanta, September 1981, 4–14.

**Friedman, Milton,** "The Marshallian Demand Curve," *Journal of Political Economy*, December 1949, *57*, 463–95

______, *Price Theory*, Chicago: Aldine Publishing Company, 1976, ch. 3.

**Gramley, Lyle E.,** Speech before community leaders in Seattle, Washington, September 11, 1980; reprinted by Federal Reserve Bank of Dallas, *Voice*, November 1980.

**Hausman, Jerry,** "Income and Payroll Tax Policy and Labor Supply," in *The Supply-Side Effects of Economic Policy*, Proceedings of the 1980 conference sponsored by the Center for the Study of American Business and the Federal Reserve Bank of St. Louis, May 1981, 175–202.

**Keeley, Michael C.,** *Labor Supply and Public Policy: A Critical Review*, New York: Academic Press, 1981.

**Tax Foundation,** *Facts and Figures on Government Finance, 1981*; New York: The Foundation, 1982.

**U.S. Council of Economic Advisers,** *Economic Report of the President*, Washington: USGPO, 1980.

# Relative Risk Aversion in Comparative Statics

*By* ELIAKIM KATZ*

In his seminal paper on the theory of the competitive firm under price uncertainty, Agnar Sandmo (1971) claims that an increase in a proportional profits tax (with full loss offset) will reduce, leave unchanged or increase the firm's output accordingly as relative risk aversion $R$ is increasing, constant or decreasing. Unfortunately, this tidy result is founded upon an error.[1] It is the purpose of this note to point out the error and correct it; thus showing what *can* be said about the effect of a profits tax on the firm's output.

The error arises because Sandmo uses a measure of $R$ which is profits based; that is, he defines

$$R = -\Pi u''(\Pi)/u'(\Pi), \tag{1}$$

where $\Pi$ is the firm's net profits, $u(\ )$ is the utility of profits, $u' = \partial u/\partial \Pi$, etc.

This definition of $R$ leads to two problems. First, it contains an internal contradiction rendering meaningless the assumptions of constant or a decreasing relative risk aversion. To see this, recall that Sandmo's analysis deals with a firm which in certain states of the world may face negative profits and compare $R$ as defined in (1) when profits are negative and when they are positive. Clearly, $R$ associated with the negative profits will be negative and hence smaller than the positive $R$ associated with positive profits. Hence, the very definition of $R$ as given in (1) is incompatible with decreasing or constant relative risk aversion.

The second problem with (1) (and this indicates the way out of the paradox described above) is that measures of risk aversion are properly defined upon wealth $W$ and not upon profits;[2] that is,

$$R = -Wu''(W)/u'(W) \tag{2}$$

is the correct definition. Once this is recognized, then since wealth cannot in general be negative, the paradox disappears.

Now, letting $W_0$ be the firm's initial wealth, $x$ the firm's output, $c(x)+B$ the firm's costs, $p$ the (random) output price, and $t$ the profits tax rate, the firm's final wealth $W$ is given by

$$W = W_0 + (px - c(x) - B)(1-t). \tag{3}$$

Assuming that the firm maximizes $Eu(W)$ with respect to $x$, we know by adapting Sandmo's result that the effect of an increase in $t$ on $x$ will have the same sign as

$$\phi(W) = -Eu''(W)(p - c'(x)) \times (W - W_0)/(1-t) \tag{4}$$

which, using the (correct) definitions of $R$ and absolute risk aversion $A$, may be written

$$\begin{aligned} \phi(W) &= ER(W)u'(W)(p - c'(x))/(1-t) \\ &\quad - EA(W)u'(W)(p - c'(x))W_0/(1-t). \end{aligned} \tag{5}$$

Now, from the work of Sandmo and of Peter Diamond and Joseph Stiglitz (1974), we know that the first term on the right-hand side of (5) is negative, zero or positive according as $R$ is decreasing, constant or increasing and that the second term on the right-hand side of (5) is negative, zero, or positive according as $A$ is increasing, constant, or decreasing.

---

*Department of economics, Bar-Ilan University, Israel.

[1] This error has been perpetuated by others with respect to other organizational forms. See, for example, John Hey (1981, p. 370).

[2] See Kenneth Arrow (1965), and John Pratt (1964).

Hence, the following table summarizes what can be said about the sign of $\phi$, that is, about the effect of an increase in profits tax on output.

| | *R* Increasing | *R* Constant | *R* Decreasing |
|---|---|---|---|
| *A* Increasing | $\phi$? | nonfeasible | nonfeasible |
| *A* Constant | $\phi > 0$ | nonfeasible | nonfeasible |
| *A* Decreasing | $\phi > 0$ | $\phi > 0$ | $\phi$? |

## REFERENCES

**Arrow, Kenneth J.**, *Aspects of the Theory of Risk Bearing*, Helsinki, 1965.

**Diamond, Peter A. and Stiglitz, Joseph E.**, "Increases in Risk and in Risk Aversion," *Journal of Economic Theory*, July 1974, *8*, 337–60.

**Hey, John D.**, "A Unified Theory of the Behavior of Profit-Maximizing, Labor-Managed and Joint-Stock Firms Operating under Uncertainty," *Economic Journal*, June 1981, *91*, 364–74.

**Pratt, John W.**, "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, *32*, 122–36.

**Sandmo, Agnar**, "On the theory of the Competitive Firm Under Price Uncertainty," *American Economic Review*, March 1971, *61*, 65–73.

# Existence of Consistent Conjectures: Comment

By Arthur J. Robson*

In an interesting recent contribution to this *Review*, Timothy Bresnahan defines the notion of a *consistent conjectural equilibrium* for duopoly. (See also the references cited by Bresnahan for related work.) Within the class of conjectural variation equilibria (as in A. L. Bowley, 1924), equilibria with the following property are sought: The *conjectural variation function* and the derived *reaction function* should be locally identical for each of the two firms. To paraphrase William Fellner (1949), firms are then "right for the right reason." This concept of a consistent conjectural equilibrium (*CCE*) has inherent interest and holds out the prospect of reducing the "embarrassment of riches" resulting from the great multiplicity of conjectural variation equilibria.

Bresnahan is able to find a *CCE* when demands and marginal costs are *linear*. Furthermore, he shows that this *CCE* is *unique* within the class of *polynomials*. This suggests two questions: 1) Surely it is desirable to use, at least, the class of analytic functions rather than polynomials? (That is, can uniqueness be established within this richer class of functions?) 2) Surely Bresnahan's results ought to be robust to generalization of demand and cost functions? (That is, does a unique (analytic) *CCE* continue to *exist* with nonlinear demands and/or marginal costs?)

It is necessary, first of all, to reformulate Bresnahan's problem. This is shown to involve the solution of an (ominously) unusual type of differential-difference equation. Eventually, a positive answer can be obtained for the first question. That is, Bresnahan's *CCE* appears to be unique within the class of *analytic* functions and not merely within the class of polynomials. Of much greater importance, however, the technique developed here also suggests a negative answer to the second question. That is, simple examples are readily found for which existence of an analytic *CCE* is problematic.

## I. Consistent Conjectural Equilibria—Reformulation, Possible Nonexistence

A simple example which permits the main points to be made is as follows. Suppose that average costs are constant and identical for the two firms and hence, without loss of generality, zero. Suppose that firm 1's inverse demand and profits are given by

$$(1)\quad P_1(q_1,q_2)=a-(q_1+q_2)-bq_1/2-cq_1^2/3,$$

$$\pi_1(q_1,q_2)=P_1(q_1,q_2)q_1.$$

Symmetrically for firm 2. Each firm maintains a conjecture concerning its rival's behavior. (Suppose initially that these conjectures are distinct for the sake of expositional clarity. Eventually symmetry will be imposed for the sake of simplicity.) Firm 1's conjecture about firm 2 is

$$(2)\qquad q_2=\alpha_2+\psi_2(q_1),$$

where $\alpha_2$ is parametric for firm 1. Similarly, firm 2's conjecture concerning firm 1 is

$$(3)\qquad q_1=\alpha_1+\psi_1(q_2).$$

It is supposed that firm 1 explicitly chooses the parameter $\alpha_1$ (rather than $q_1$ directly) and, similarly, firm 2 chooses $\alpha_2$. It is assumed that choice of $(\alpha_1,\alpha_2)$ yields a unique payoff combination $(\pi_1,\pi_2)$. Consider then the profit-maximizing behavior of firm 1. The first-order condition is

$$(4)\quad a-2q_1-(\alpha_2+\psi_2(q_1))-q_1\psi_2'(q_1)-bq_1-cq_1^2=0,$$

which thus describes firm 1's best-reply function for $q_1$ in terms of $\alpha_2$. The implied relationship between $q_1$ and $q_2$ can be obtained by substituting for $\alpha_2$ in (4) from (2). Then (4) becomes

---

*Hoover Institution, Stanford University, and University of Western Ontario. I thank Tim Bresnahan for helpful comments.

$$(5)\quad a-2q_1-q_2-q_1\psi_2'(q_1)-bq_1-cq_1^2=0.$$

The second-order condition for firm 1's maximization problem is

$$(6)\quad (2+b+2cq_1)+2\psi_2'(q_1)+q_1\psi_2''(q_1)\geqslant 0,$$

which must hold identically throughout some neighborhood.

An entirely parallel argument describes firm 2's best-reply to firm 1. It is convenient at this stage to recognize that $\alpha_1$ and $\alpha_2$ may now be set equal to zero, if the conjectures $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are those that would apply in equilibrium. Consistency as defined by Bresnahan requires that the relationship between $q_1$ and $q_2$ describing firm 1's best-reply to firm 2, as in (5), should be locally identically the conjecture held by firm 2 concerning firm 1's behavior, as in (3), with $\alpha_1=0$. Hence (5) becomes the requirement that

$$(7)\quad a-2\psi_1(q_2)-q_2-\psi_1(q_2)\psi_2'(\psi_1(q_2)) - b\psi_1(q_2)-c(\psi_1(q_2))^2=0,$$

which is to hold as an identity in $q_2$. It is assumed that an equilibrium in the usual Bowley sense exists so that

$$(8)\quad q_1^*=\psi_1(q_2^*);\ q_2^*=\psi_2(q_1^*).$$

It should be noted that (7) is an integrated form of Bresnahan's comparable equation defining a *CCE*. Equation (7) incorporates the first-order condition for profit maximization as well as the requirement of consistency.

Consider the possibility that $\psi_1'(q_2^*)=0$. Differentiation of (7) at $q_2^*$ readily yields a contradiction. Hence (7) may be rewritten, at least in a neighborhood of $q_2^*$, as

$$(9)\quad q_1\psi_2'(q_1)=a-2q_1-\psi_1^{-1}(q_1)-bq_1-cq_1^2.$$

The final requirement imposed is that $(q_1^*,q_2^*)$ belong to the "economically plausible" region. This can be obtained, for example, by requiring stability of the following simple associated dynamic system:

$$(10)\quad \begin{aligned} q_{1,t}&=\psi_1(q_{2,t-1})=\psi_1(\psi_2(q_{1,t-2}));\\ q_{2,t}&=\psi_2(q_{1,t-1})=\psi_2(\psi_1(q_{2,t-2})).\end{aligned}$$

For global or local asymptotic stability at $(q_1^*,q_2^*)$, it is necessary that

$$(11)\quad |\psi_1'(q_2^*)\psi_2'(q_1^*)|\leqslant 1.$$

Suppose now symmetry is finally imposed on the conjectures. Now (9), (8), (6), and (11) become, respectively,

$$(12a)\quad q\psi'(q)=a-2q-\psi^{-1}(q)-bq-cq^2$$

$$(12b)\quad q^*=\psi(q^*)$$

$$(12c)\quad (2+b+2cq)+2\psi'(q)+q\psi''(q)\geqslant 0$$

$$(12d)\quad |\psi'(q^*)|\leqslant 1,$$

which are necessary conditions for any *CCE* which might exist.

The basic equation here is (12a) which is a mixed differential-difference equation in which the difference is endogenous. Although equations answering this general description are treated in the mathematical literature, this particular case has not been investigated and does not appear to correspond to a well-posed physical problem. (See L. El'sgol'ts and S. B. Norkin, 1973, for the results apparently available.)

Equations (12) can be simplified somewhat. Differentiation of (12a) and use of (12b) and (12c) implies that either $\psi'(q^*)=1$ or $\psi'(q^*)\in[-1,0)$. In the first case, it can be shown that $\psi''(q^*)<0$ if $c$ satisfies (13d) below, and that two-sided stability cannot then hold. Hence (12) reduces to

$$(13a)\quad q\psi'(q)=a-2q-\psi^{-1}(q)-bq-cq^2$$

$$(13b)\quad q^*=\psi(q^*)$$

$$(13c)\quad \psi'(q^*)\in[-1,0)$$

$$(13d)\quad c>-(4+b)/2a.$$

The basic result is

THEOREM 1: *Equation* (13a) *provides a recursive rule for finding all derivatives of* $\psi(q)$ *at* $q^*$ *for any* $q^*$ *satisfying* (13c). *However, the associated Taylor series will generally diverge for all* $q^*$ *such that* $\psi'(q^*) \in (-1,0)$ *unless* $\psi(q)$ *is linear.*

PROOF: (See the Appendix.)

This theorem has the following two immediate corollaries.

COROLLARY 1: *If* $b>0$, $c=0$, *this case was treated by Bresnahan. The unique linear solution of* (13) *is easily seen to be*

$$(14)\quad \psi(q)=\left[-1+\tfrac{1}{2}\left(\sqrt{b^2+4b}-1\right)\right][q-a],$$

*which satisfies* $\psi'(q^*) \in (-1,0)$. *This then seems to be the unique analytic solution of* (13), *for* $\psi'(q^*) \in (-1,0)$.

This serves to exemplify the sense in which the first question posed in the introduction has an affirmative answer. However,

COROLLARY 2: *If* $b>0$, $c \neq 0$ *subject to* (13d), *no matter how small c may be, it is easily seen that* (13) *has no linear solution.* (*Indeed, it has no polynomial solution.*) *Hence no analytic solution generally exists either for* $\psi'(q^*) \in (-1,0)$.

In this sense, the second question posed above is answered in the negative, and the existence of the *CCE* found by Bresnahan must be viewed as something of an accident.

## APPENDIX: PROOF OF THEOREM 1

Differentiation of (13a) yields

$$(\text{A1})\quad \psi''(q)q+\psi'(q) = -2-1/\psi'(r)-b-2cq,$$

where $r=\psi^{-1}(q)$. Setting $q=q^*$ yields

$$(\text{A2})\quad \psi''(q^*)q^*+\psi'(q^*) = -2-1/\psi'(q^*)-b-2cq^*.$$

Continuing this process, it is easily shown that, in general,

$$(\text{A3})\quad \psi^{(n+1)}(q^*)q^*=\frac{\psi^{(n)}(q^*)}{(\psi'(q^*))^{n+1}} - \psi^{(n)}(q^*)+\text{terms in } \psi'(q^*),\dots,\psi^{(n-1)}(q^*),$$

which provides a recursive formula for finding all the derivatives of $\psi(q)$ at $q^*$. Suppose that $x=1/\psi'(q^*)<-1$, if $\psi'(q^*) \in (-1,0)$. Now (A3) can be written as

$$(\text{A4})\quad \psi^{(n+1)}(q^*)q^*=x^{n+1}\psi^{(n)}(q^*)+R,$$

where $R$ is other terms of lower order in $x$.

Consider then the (formal) Taylor series for $\psi(q)$:

$$(\text{A5})\quad \sum_{n=0}^{\infty}\frac{\psi^{(n)}(q^*)(q-q^*)^n}{n!}.$$

However, this series diverges by appropriate ratio-test (see W. Rudin, 1974) since the

$$(\text{A6})\quad \limsup_{n\to\infty}\left|\frac{\psi^{(n+1)}(q^*)(q-q^*)}{(n+1)\psi^{(n)}(q^*)}\right|=\infty,$$

for all $q \neq q^*$, using (A4). The only general exceptions require $\psi(q)$ to be a polynomial so that $\psi^{(n)}(q^*)=0\ \forall n \geqslant N$. But if $\psi(q)$ is a polynomial, (13a) requires $\psi^{-1}(q)$ to be one also. This is only possible if $\psi(q)$ is *linear*.

## REFERENCES

**Bowley, A. L.**, *The Mathematical Groundwork of Economics*, Oxford: Oxford University Press, 1924.

**Bresnahan, T. F.**, "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, *71*, 934–45.

**El'sgol'ts, L. E. and Norkin, S. B.**, *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, New York: Academic Press, 1973.

**Fellner, William J.**, *Competition Among the Few*, New York: Knopf, 1949.

**Rudin, W.**, *Real and Complex Analysis*, New York: McGraw-Hill, 1974.

# Existence of Consistent Conjectures: Reply

*By* TIMOTHY F. BRESNAHAN*

Arthur Robson's excellent comment introduces a new formulation of the problem of consistent conjectures. My reply uses the new formulation in order to resolve Robson's troubling example of nonexistence, and to make the problem clearer. It also shows why my original approach was sensible only for the linear demand curve case. A slightly more general approach used here has consistent conjectures in Robson's quadratic inverse demand example.

## I. Robson's Reformulation: An Extension

Robson writes the consistent conjectures problem as selection of the strategic variables for oligopoly. He introduces the notation $\alpha$ for strategic variables in general, and considers a broad class of games that are Nash equilibrium with $\alpha$ as strategic variables. Each member of this class is specified by allowing quantity to be a specific function of $\alpha$:

$$(1)\quad q_1 = q_1(\alpha_1, \alpha_2);\; q_2 = q_2(\alpha_1, \alpha_2),$$

which I abbreviate $q = q(\alpha)$. The payoffs can easily be written in terms of the $\alpha$'s.

$$(2)\quad \Pi_i(\alpha) = \Pi_i(q(\alpha)) = D_i^{-1}(q(\alpha))\,q_i(\alpha) - (q_i(\alpha)),$$

where $P_i = D_i^{-1}(q)$ is the inverse demand curve.

Then Nash equilibrium for the new game so defined is determined by

$$(3)\quad 0 = \frac{\partial \Pi_i(\alpha)}{\partial \alpha_i} = \frac{\partial \Pi_i}{\partial q_1}\frac{\partial q_1}{\partial \alpha_i} + \frac{\partial \Pi_i}{\partial q_2}\frac{\partial q_2}{\partial \alpha_i}; \qquad i = 1,2.$$

From the $\alpha$'s that solve (3), we can calculate equilibrium quantities and prices. This formulation includes the usual oligopoly solution concepts by choice of particular changes of variables $q(\alpha)$. The identity $q_i = \alpha_i$ yields Cournot (quantity) equilibrium, while changing along the demand curve $q = D(\alpha)$ yields Bertrand (price) equilibrium.

The first advantage of this formulation is that it makes absolutely clear the relationship of "conjectures" to the theory of noncooperative games; that is, every equilibrium studied is a Nash equilibrium in some strategic variables. The class of strategic variables is wide enough to include both prices and quantities. The second advantage of this formulation is that it makes definition of "consistency" very simple mathematically. A consistent conjectures equilibrium is a change of variables $q(\alpha)$ such that the reaction functions $\alpha_i(\alpha_j)$ are constant. Since the reaction functions are defined by (3), this is simply the condition that

$$(4)\quad \partial^2 \Pi_i(\alpha)/\partial \alpha_i \partial \alpha_j = 0; \quad i = 1,2, \quad j \neq i.$$

This is a pair of nonlinear partial differential equations for the functions $q(\alpha)$.

In this notation, the limitation of my original approach was to consider only changes of variables that had $\partial q_i/\partial \alpha_i = 1$, although $\partial q_j/\partial \alpha_i$ was allowed to vary.

## II. Quadratic Inverse Demand

Robson showed that consistent conjectures as earlier defined do not exist for duopolists with zero costs and the quadratic inverse demand system:

$$(5)\quad P_i = a - \frac{b}{2}q_i - (q_i + q_j) - \frac{c}{3}q_i^2; \qquad i = 1,2.$$

This example brings out the shortcomings of my earlier approach. Inverting the system (5) in order to compute the direct demand system involves solving a quartic equation. It is not surprising, therefore, that my implicit assumption $\partial q_i/\partial \alpha_i = 1$ leads to difficulties. This assumption rules out Bertrand equilibrium, among others.

---

*Department of economics, Stanford University. I had many stimulating conversations with Arthur Robson, Stephen Turnbull, and Peter Huang.

A change of variables $q(\alpha)$ can be found that leads to consistency in this case. In fact, the change takes the form $q_i = f(\alpha_i) + g(\alpha_j)$. A technical appendix (available upon request from the author) shows that any functions $f(\ )$ and $g(\ )$ and constants $K_1, K_2, \lambda$ that satisfy

$$(6) \qquad cf^2(t) - \lambda f(t) = K_1 - g(t)\ \forall t;$$

$$cg^2(t) + [\lambda + 2 + b]g(t) = K_2 - f(t)\ \forall t.$$

can, by equation (4), be shown to yield consistent conjectures. In general, the solution to the equations (6) involves solving a quartic not unlike the demand system. In the case of linear demand ($c = 0$) we get

$$(7) \quad g'(t)/f'(t) = \left[ -(2+b) + \sqrt{(2+b)^2 - 4} \right] \Big/ 2.$$

which could also be found using (A9) in my original paper.

Thus a natural reformulation of the problem suggested by Robson's comment reverses his finding of nonexistence. The broader class of solutions found here reduces to my earlier result when demand is linear.

## III. Conclusion

The task of a consistency criterion is to select one noncooperative oligopoly solution concept from the many possible. The selection criterion should always choose something (existence), yet should never choose many things (uniqueness). In this reply, the criterion was weakened in order to ensure existence; whether this destroys uniqueness is an open question.

## REFERENCES

**Bresnahan, T. F.**, "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, *71*, 934–45.

**Robson, A. J.**, "Existence of Consistent Conjectures: Comment," *American Economic Review*, June 1983, *73*, 454–56.

# The Economics of Superstars: Comment

By PETER BOWBRICK*

Sherwin Rosen's interesting analysis in a recent issue of this *Review* goes a long way to explain the star phenomenon, but it misses two reasons why, though most industries have stars and superstars, some have only superstars, or have superstars and a few insipid stars. This may happen when several stars combine to make a product, or when the product is consumed by several people at the same time.

The key may be found in David Ogilvy's report of his work with Gallup in the 1930's: "I discovered that some stars had a *negative* effect at the box office; their names on the marquee repelled more ticket buyers than they attracted. The list, which I called Box Office Poison and classified TOP SECRET, included some of the most famous names in show business, and ruined their careers" (p. 49).

One cannot treat this repulsiveness just as negative attractiveness. Two popular stars, each hated by a fair number of people, but not the same people, would guarantee a flop —if the Magnificent Seven had each been hated by a different 14.28 percent of the population, there would have been no audience. Clearly, most actors must be bland, neither loved or hated, as a film could afford to have only one or two stars who were sufficiently interesting to be hated. Instead of the galaxy of stars of the 1930's, we have the superstars of today: those who are interesting, but have few haters.

Also important is the fact that people tend to go to view films in couples or family groups, so one hater lost the film two or more customers, an effect that was particularly strong when nonhaters saw other films as close substitutes. The advent of portable radio switched the audience from the family sitting round the set to the individual or to a homogeneous group, so even widely disliked stars could succeed with specialized audiences. Sid Vicious could be a superstar where previously the broad appeal of a Crosby had been necessary.

The phenomenon is not confined to people. Food marketing in particular shows the importance of haters. Complex flavors and strong flavors increase the chance that someone will dislike the product, so it is safer to market the bland (especially with ready-to-eat foods). Similarly, the fact that a meal is eaten by a whole family means that a single hater can change the family shopping list.

The model is strongest when consumers have a choice strategy that rejects a product if it has attribute $X$ (for example, garlic in the stew), or that judges a product by its worst attribute. It is weaker if the strategy is additive, with good attributes offsetting bad. It will reinforce some of the superstar cases mentioned by Rosen.

If there were no aversion, a single superstar would still attract in spite of a poor supporting cast, and his fans would bring along others indifferent to his charms. If, in addition, there was additive evaluation, one superstar would have the same attraction as three stars, or as five starlets. Evaluation processes must therefore be taken into account. The model is particularly useful in explaining the blandness of commercial television programs and of mass-produced food.

## REFERENCES

**Rosen, Sherwin,** "The Economics of Superstars," *American Economic Review*, December 1981, *71*, 845–58.

**Ogilvy, David M.,** "*Blood, Brains and Beer*," London: Hamish Hamilton, 1978.

---

*Institute of Agricultural Economics, Oxford University.

# The Economics of Superstars: Reply

*By* SHERWIN ROSEN*

Peter Bowbrick raises some interesting issues that deserve elaboration. Consider a differentiated product with attribute $q$ that sells for $p(q)$ per unit in the market. Assume an economy of scale in consumption of a particular type of good among a small group of buyers, a family, for example. Positive consumption of that variety requires incurring a fixed set-up cost $k$, independent of the quantity consumed. The fixed cost is incurred every time a member chooses an independent variety. For example, the location of consumption may be away from home. Then $k$ represents the cost of transport in an auto with excess capacity. More generally, $k$ derives from scale economies in home production, for which the good in question is intermediate product. Several types of outcomes are possible. Each member can go it alone and make a purely private consumption decision. This has the virtue of allowing members to consume their most preferred varieties, but entails duplication of set-up costs. Alternatively, the group may make a collective decision and choose a single variety that is consumed by all members.[1] The group decision is equivalent to choosing a public good. The variety actually chosen compromises among the most preferred choices of each person, but exploits the scale economy and allows greater consumption of other goods as compensation.

Some possibilities are illustrated in Figure 1 for a group of size 2. Bid-price functions (price-attribute indifference curves) for both members are shown by the curves labeled $\theta^i$ (the utility direction is southeast). The opportunity locus is $k + p(q)$ when consumption decisions are made independently. The optimum good for each person is characterized by equality between marginal rates of substitution between $q$ and other goods and the incremental cost of $q$ for each person: $\partial\theta^i/\partial q = p'(q)$, $i = 1,2$ (points $a_1$ and $a_2$ in Figure 1). Alternatively, the per capita opportunity locus becomes $(k/2) + p(q)$ when members choose a single variety and share fixed costs. The figure illustrates a case where a public decision Pareto dominates individualistic ones. This happens whenever the per capita group constraint cuts through the shaded area in the figure, the locus of possible Pareto-improving group allocations relative to independent decisions at $a_1$ and $a_2$. All points between $b_1$ and $b_2$ raise utility of at least one member without lowering utility of the other. All these points potentially satisfy a Samuelsonian public goods condition, $(1/n)\Sigma_i\,\partial\theta^i/\partial q = p'(q)$, depending on welfare weights among group members. For example, both persons achieve larger utility at point $c$ than at their most-preferred independent locations. Cost savings due to exploiting group scale economies more than compensate for compromising on a variety somewhere in the middle.

More structure is needed to nail the precise location of the compromise, say, a Nash bargaining game or an effectively altruistic head. The resolved preference structure might replace my original specification of preferences, but the essential argument regarding concentration and skew of market shares and reward among sellers is otherwise not affected. Nevertheless, the point is clear that small collective decisions of this sort promote a tendency toward mediocrity of choices within groups, in the literal sense of the median. This illustrates one aspect of the "blandness" of goods to which Bowbrick refers, but is not necessarily assured, because group decisions need not dominate individualistic ones. Experimentation with the figure reveals that a public goods decision is more likely: (i) the larger the scale economy; (ii) the greater the uniformity of preferences;

---

*University of Chicago and National Opinion Research Center. I am indebted to the National Science Foundation for financial support.

[1] Coalitions of various subgroups should be considered, but space does not permit that development here.

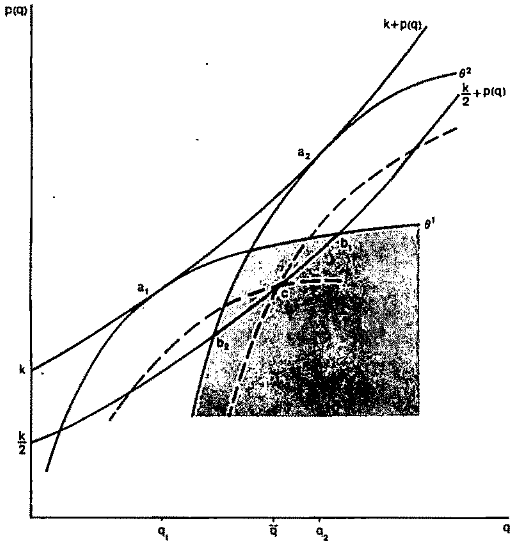


FIGURE 1

and (iii) the greater the degree of substitution between these goods and other goods.

That collective decisions within households reduce variety among members has no additional relevance for the superstar phenomenon unless there are scale economies for sellers. Profitability and rents are only weakly related to firm or personal market size in the absence of scale economies in market production; and the skewed and concentrated market share and reward distributions I attempted to explain do not arise, whether or not there are collective decisions within families. Market scale economies are necessary for superstars even if firm's outputs are jointly produced (multiple actors).

The broader issues raised by Bowbrick's remarks refer to which varieties appear on the market, and are related to Harold Hotelling's theory of spatial competition. Scale economies in market production remain necessary for observing large concentrations of market shares and rewards and for focusing the varieties actually offered toward median preferences. A reinterpretation of Figure 1 illustrates the point. Now think of a continuous distribution of preferences among consuming units in the market as a whole. Tangencies along the upper opportunity locus represent the spectrum of equilibria when everyone independently consumes their most preferred variety. This would be the market solution were there no scale economies in production. If a seller can exploit scale economies by massing customers with varying preferences on a particular variety, the good can be sold on better terms. For example, a seller who could configure a variety and price at the intersection of $\theta^1$ and $\theta^2$ would attract all consumers with intermediate preferences, and unit costs of production would fall. If the price were lower that variety would attract consumers whose preferences are even more extreme than those of $\theta^1$ and $\theta^2$, and unit costs would fall even more. Scale economies in market production therefore push the market to reach compromises among buyers that resemble a local public goods equilibrium.[2] We expect to see firms concentrating some of their competitive energies on designing goods to cater to "mass tastes" under these circumstances. Persons with minority preferences might not be served very well or at all.

There are many real problems where considerations such as these are relevant. A single provider of television programs must compromise between families as well as within them. The same is true of subscription schedules of major symphony orchestras, where works of modern day composers are seldom played because they cater to minority tastes and offend the majority. Cases such as these involve large elements of public goods for the market as a whole. More competition, if feasible, would provide greater variety. This is clear enough in contrasting most print media with television. Perhaps production economies are smaller for the former, but licensing restrictions probably contribute to these differences as well. Surely cable television ultimately will increase program variety and alter rents to stars in that activity. Greater variety itself reduces the size of each submarket, which works to reduce rents; but better matching of tastes with variety increases both willingness to pay and attracts additional consumers who previously did not purchase, which works to increase them.

---

[2] This is a spatial model of monopolistic competition. See Kelvin Lancaster (1975).

There remains intuitive plausibility to the idea that some products and services must be "degraded" to appeal to large numbers of buyers. Yet one can take comfort only in introspection on that score and none whatsoever in economic theory, which admits no such concepts: neoclassical value theory is value free! I finessed this issue in my paper by defining markets very narrowly and by maintaining a specification that preserved pure competition rather than monopolistic competition. Competition among sellers implied a unique market equilibrium price of services. Degradation of a seller's product occurred through congestion and external diseconomies, with relatively more "crowding out" and larger market shares of better sellers in equilibrium. This specification was technically necessary for limiting scale economies and achieving a nondegenerate distribution of sellers, while maintaining pure competition.[3] A monopolistically competitive formulation limits scale economies in another way, namely, by a seller losing buyers to competitors at his market fringes. My specification has the effect of putting the distinction between, say, serious novels and trashy ones on the same footing as that between legal services and baseball—entirely different markets. Why writers of trash have larger audiences than serious novelists is better argued over lunch than in print, but the thought that it might be as meaningful a categorization on the supply side as that between economists and rock musicians is perhaps not entirely implausible. The fact that the highly talented take home most of the marbles within these narrow categories can be explained within the standard competitive framework, but a monopolistically competitive model undoubtedly would improve our understanding of this problem even more.

## REFERENCES

**Bowbrick, Peter,** "The Economics of Superstars: Comment," *American Economic Review*, June 1983, *73*, 459.

**Hotelling, Harold,** "Stability in Competition," *Economic Journal*, March 1929, *34*, 41–57.

**Lancaster, Kelvin,** "Socially Optimal Product Differentiation," *American Economic Review*, September 1975, *65*, 567–85.

**Rosen, S.,** "The Economics of Superstars," *American Economic Review*, December 1981, *71*, 845–58.

——, "Authority, Control and the Distribution of Earnings," *Bell Journal of Economics*, Autumn 1982, *13*, 311–23.

---

[3] I have extended the basic problem to firm size and managerial reward distributions across ranks in hierarchical firms in my 1982 article.

# Stock Returns, Real Activity, Inflation, and Money: Comment

*By* RATI RAM AND DAVID E. SPENCER*

An interesting empirical phenomenon which has received considerable attention is the observed negative correlation between real stock returns and inflation (both expected and unexpected components). In an interesting recent article in this *Review*, (1981), Eugene Fama sought to explain this phenomenon. His explanation is broadly on the following lines: contrary to the Phillips curve hypothesis, there is a negative correlation between inflation and real economic activity; on the other hand, there is a positive association between real activity and stock returns. The observed negative correlation between real stock returns and inflation is thus due to the conjunction of these two relationships, or due to what he calls the "proxy effect."

Fama's procedure consists essentially of three main steps. First, estimate two sets of inflation regressions, one based on a Fisher-type equation in the rational expectations framework, and the other derived from a particular form of the money demand equation. The purpose of these is to show a negative association between inflation and real activity. Second, estimate certain capital expenditure functions and real stock return regressions primarily to show that stock returns and real activity are positively associated. Third, estimate real stock return regressions in which both inflation and real activity variables are included, the main objective being to show that holding real activity constant, inflation (at least expected inflation) is not negatively correlated with stock returns. However, since that objective is not quite met when inflation and real activity variables are used as regressors in the stock return functions, growth of money is added as another "explanatory" variable to show an attenuation of the negative effect of inflation on stock returns. Even that strategy succeeds only partially, apart from the difficulty of justifying money growth variables in these real stock return regressions.

The evidence adduced by Fama has the apparent consequence of calling into question conventional macroeconomic theories of the Phillips curve which suggest a *positive* relationship between inflation and real activity. The purpose of this comment is to take issue with that conclusion, and thus with Fama's explanation of the relationship between real stock returns and inflation. We find considerable empirical evidence which is inconsistent with the world described by Fama, but generally consistent with conventional macroeconomic theory.

Our alternative explanation essentially follows from the fact that we employ a different inflation equation, derived from a Fisher-Phillips relationship, and our use of different but equally plausible variables to represent real activity. We also estimate stock return equations of much the same character as Fama's. Our evidence suggests an explanation which is essentially the reverse of Fama's: the negative association between stock returns and inflation can be postulated to arise from a conjunction of a *positive* association between inflation and real activity and a negative association between real activity and stock returns. These results along with others suggest that the observed negative correlation (between stock returns and inflation) may be best explained as a straightforward reflection of the Mundell-Tobin hypothesis rather than by Fama's indirect explanation. It should be noted that while neither Fama's nor our regression functions for inflation and stock returns are derived from a specific general equilibrium model, our procedure (besides reflecting "prevailing wisdom") seems conceptually at least as rigorous and plausible as Fama's, and our explanation fits the data at least as well.

---

*Professor of economics, Illinois State University, and associate professor of economics, Washington State University, respectively.

The rest of this paper is structured as follows. In Section I, we reexamine the relationship between inflation and real economic activity. Fama's procedure is discussed and an alternative procedure explained and employed. The next section reexamines the relationship between real stock returns and real economic activity using several alternative measures of real activity. Section III briefly discusses the regression results and also reports the results of a test of the direction of "causality" between inflation and stock returns.

## I. Inflation and Real Activity

To find a negative correlation between inflation and real economic activity, Fama relies on a transformed money demand equation with some special features. A "typical" empirical money demand equation (see David Laidler) might be given by

$$\ln M_t = b_0 + b_1 \ln A_t + b_2 \ln R_t + b_3 \ln P_t + u_t, \tag{1}$$

where $M_t$ represents nominal money balances, $A_t$ is an appropriate measure of real economic activity, $R_t$ is a nominal rate of interest, $P_t$ is the price level, and $u_t$ is a random disturbance term; all for period $t$. It is frequently assumed that $b_3 = 1.0$ although that assumption has recently been criticized (see, for example, Y. P. Mehra). Fama apparently accepts such a specification, but in order to use money demand theory to explain inflation he takes first differences of equation (1) and, assuming that $A_t$ and $R_t$ are exogenous and that $b_3 = 1.0$, solves for $\Delta \ln P_t$, the rate of inflation. Such a procedure yields

$$\Delta \ln P_t = - b_1 \Delta \ln A_t - b_2 \Delta \ln R_t + b_4 \Delta \ln M_t + \varepsilon_t, \tag{2}$$

where $b_4 = 1.0$, $\varepsilon_t = u_t - u_{t-1}$, and $\Delta \ln P_t$ equals $\ln P_t - \ln P_{t-1}$ (and similarly for other variables). However, this is not quite the equation Fama estimates. He makes two changes, the first of which is to include a constant term in equation (2). Such an addition of an irrelevant variable is of no great consequence except that in Fama's regressions (reported in his Table 1), the constant term is always statistically significant, which may be indicative of some kind of specification error. The second change is the exclusion of an interest rate variable with the justification that "the general finding in the money demand literature [is] that the interest rate is the weakest variable in empirical money demand equations" (p. 549). While there may be a sense in which such a statement is true, virtually all recent empirical money demand studies indicate the importance of a nominal interest rate variable in a properly specified money demand equation. Laidler states: "Thus the importance of the rate of interest for the demand for money is now established beyond any reasonable doubt..." (p. 120). Thus, there is additional cause for concern that Fama's regression equations are misspecified.[1]

An additional problem arises from the distribution of $\varepsilon_t$, the disturbance in equation (2). Unless the disturbance in equation (1), $u_t$, follows a first-order autoregressive structure with a unitary root, $\varepsilon_t$ will be serially correlated and the usual test statistics obtained from ordinary least squares (*OLS*) will be invalid. (See C. I. Plosser and G. William Schwert for a recent discussion of this issue.) Prevailing empirical evidence appears to suggest that $u_t$ can be reasonably represented as a first-order autoregressive process, but with an autoregressive parameter considerably less than unity.

For these reasons, we are inclined to question the usefulness of Fama's inflation regressions derived from money demand relations for investigating the relationship between inflation and real activity. Fortunately, there is another framework which avoids these problems and in which this same relationship can be examined. Following the approach used by Vito Tanzi, the familiar

---

[1] In a recent paper, Thomas Cooley and Stephen LeRoy have questioned the "objectivity" of the reported negative interest elasticities in money demand functions. However, it is *not* their position that an interest rate variable does not belong in such functions, but apparently that identification issues in money demand estimation have been overlooked.

Fisher equation can be augmented in a straightforward way in order to derive an appropriate regression equation.[2]

The basic Fisher equation is given by

$$R_t = \tilde{r}_t + \tilde{I}_t, \tag{3}$$

where $R_t$ is the nominal rate of interest, $\tilde{r}_t$ is the expected real rate, and $\tilde{I}_t$ is the expected rate of inflation; all for period $t$. This equation is combined with two other hypothetical relationships. The first of these is an expectations-augmented Phillips curve which can be expressed as

$$I_t = \tilde{I}_t + \alpha A_t, \tag{4}$$

where $A_t$, as defined above, is an appropriate measure of real economic activity. Conventional theories of the Phillips curve argue that $\alpha > 0$, but Fama seems to reject these theories and argues that $\alpha < 0$ when appropriately estimated. Thus, part of the controversy has to do with the sign of $\alpha$ in (4).[3] We do not estimate (4) directly, but add the second hypothetical relationship suggested by Tanzi which indicates a positive relationship between the expected real rate of return and real activity:

$$\tilde{r}_t = r^n + \theta A_t, \tag{5}$$

where $r^n$ may be considered as the long-run (natural) real rate of interest and it is expected that $\theta > 0$.

Combining equations (3), (4), and (5), and solving for $I_t$, we obtain

$$I_t = -r^n + R_t - (\theta - \alpha)A_t, \tag{6}$$

or

$$I_t = \gamma_0 + \gamma_1 R_t + \gamma_3 A_t. \tag{7}$$

The sign of $\gamma_3$ depends, in part, on the sign of $\alpha$. If $\alpha < 0$ as Fama would seem to suggest, then $\gamma_3$ is unambiguously negative. If $\alpha > 0$, as traditionally expected, the sign of $\gamma_3$ is ambiguous. Thus we want to estimate (7) and examine the sign of the estimate of $\gamma_3$. A failure to reject the hypothesis that $\gamma_3 \geqslant 0$ is evidence against Fama's hypothesis.

Before we can estimate (7), we must choose appropriate variables for $I_t$, $R_t$, and $A_t$. For $I_t$, the rate of inflation, we use the percentage change in the Consumer Price Index. For $R_t$, we use the rate of return on Treasury bills.[4] For $A_t$, real economic activity, Fama mainly used the rate of growth of real *GNP* as well as the rate of growth of industrial production. While we do not argue that such measures are inappropriate, there are other plausible and frequently used indicators of real economic activity which might also be fruitfully considered. We have chosen to consider the following:

$E$ = employment/population ratio (expressed as deviations from the sample mean),[5]

$UNR$ = natural rate of unemployment − actual rate of unemployment,

$$GAP^6 = \frac{\text{Actual real } GNP - \text{Potential real } GNP}{\text{Potential real } GNP} \times 100$$

$GY$ = rate of growth of real *GNP* (as used by Fama).

Each of these is a positive indicator of economic activity; that is, a larger value indicates greater real activity.

In addition to examining these alternative measures of economic activity, we also estimate equation (7) using quarterly data over two sample periods, 1953:I–1972:IV and 1953:I–1978:IV, in order to see how our results might change with the inclusion (or exclusion) of the more turbulent years since 1972. These results are reported in Table 1. The first four lines of the table give estimates of (7) for each of the four measures of real activity over the shorter time period and the

---

[2] Similar augmentation of the Fisher equation has also been suggested recently by Maurice Levi and John Makin (1978, 1979) on the basis of a general equilibrium model.

[3] That is to say, do we observe a Phillips curve ($\alpha > 0$) or a "Fama curve" ($\alpha < 0$) in the U.S. economy.

[4] These variables, $I_t$ and $R_t$, are similar to those used by Fama. See the Appendix for a more detailed description of the data we use.

[5] This is an interesting variable which has been employed in a similar context in a comment on Fama (1975) by John Carlson and by Fama (1977) in his reply.

[6] This variable is similar to that used by Tanzi as an index of real economic activity.

TABLE 1—REGRESSIONS OF INFLATION ON INTEREST RATE AND REAL ECONOMIC ACTIVITY[a]

| Model | Estimated Regression | $R^2$ | *D-W* |
|---|---|---|---|
| Sample Period: | 1953:I–1972:IV | | |
| 1. | $I_t = .719 + .578R_t + 1.021E_t$<br>(1.19) (4.08) (3.99) | .56 | 2.08 |
| 2. | $I_t = -.150 + .755R_t + .427UNR_t$<br>(−.23) (4.80) (2.07) | .50 | 2.02 |
| 3. | $I_t = -.850 + .900R_t + .0767GAP_t$<br>(−1.58) (6.33) (.95) | .49 | 2.00 |
| 4. | $I_t = -.700 + .928R_t - .063GY_t$<br>(−1.24) (7.37) (−1.26) | .48 | 2.01 |
| Sample Period: | 1953:I–1978:IV | | |
| 5. | $I_t = .142 + .857R_t + 1.274E_t$<br>(.19) (4.96) (4.44) | .69 | 2.00 |
| 6. | $I_t = -2.307 + 1.419R_t - .096UNR_t$<br>(−3.19) (9.42) (−.45) | .63 | 1.98 |
| 7. | $I_t = -2.266 + 1.418R_t - .064GAP_t$<br>(−3.47) (9.88) (−.72) | .63 | 1.98 |
| 8. | $I_t = -1.757 + 1.352R_t - .062GY_t$<br>(−2.38) (9.15) (−1.10) | .63 | 1.99 |

[a] Each regression equation is estimated by the Cochrane-Orcutt iterative procedure; *D-W* indicates the Durbin-Watson statistic after the autoregressive correction; *t*-ratios are shown in parentheses; $R^2$ is the coefficient of determination in the transformed regression.

last four lines give the corresponding estimates for the full period.

In no case, either for the short period or the long period, are we able to reject the null hypothesis that $\gamma_3 \geqslant 0$; that is (at the usual levels of significance), there are no statistically significant negative coefficients on the various proxies for real economic activity. The coefficient of $E$ is significant and *positive* for both periods while that of $UNR$ is significant and *positive* for the shorter period. All other regressions yield insignificant estimates of $\gamma_3$ although $GY$ (one of Fama's choices) comes closest to satisfying Fama's hypothesis with a *t*-ratio of $-1.26$ in the short period. These results offer no evidence of a significantly negative association between inflation and real activity. The estimates that are statistically significant are all positive. The results for each regression therefore appear to be consistent with conventional theories of the Phillips curve ($\alpha > 0$) and inconsistent with Fama's position on those theories.[7]

As might be expected, a comparison of the results for each sample period suggests a deterioration of the Phillips curve relationship during the post-1972 period. Although it is *not* necessary to choose among the various measures for $A$ to reach our conclusion, it is the case that such a choice on the basis of $R^2$ would lead to the use of $E$, the only variable which consistently shows a significant positive relationship with inflation.[8]

## II. Stock Returns and Real Economic Activity

To examine the second part of his hypothesis, Fama estimates regressions of real stock returns ($RS_t$) on several measures of

---

[7] It is interesting to note that even if we take the insignificant estimates of $\gamma_3$ to be true (i.e., if we have committed Type I error), our regressions still appear to imply that $\alpha > 0$. Consider line 7 in Table 1 since the measure of real activity, $GAP$, corresponds to the measure used by Tanzi. Over a similar sample period, Tanzi found estimates of $\theta$ to be approximately 0.2 which implies an estimate of $\alpha = \gamma_3 + \theta$ which is positive (approximately 0.1).

[8] It should be noted that the lag structure of the regressors in our reported regressions is somewhat different from that in Fama's equations. He places greater emphasis on *future* values of the real activity variables. As indicated in the introduction, our interest is from the perspective of Fama's challenge to conventional macroeconomic theory. Therefore, our lag structure is appropriate for our purposes. In any event, it is the choice of real activity variable, *not* the lag structure, which plays the more important empirical role.

TABLE 2—REGRESSIONS OF REAL STOCK RETURNS ON REAL ECONOMIC ACTIVITY[a]

| Model | Estimated Regression | $R^2$ | $D$-$W$ |
|---|---|---|---|
| Sample Period: 1953:I-1972:IV | | | |
| 1. | $RS_t = -2.401 - 2.472E_t$<br>$(-5.58)\quad(-5.30)$ | .33 | 1.98 |
| 2. | $RS_t = -2.108 - 1.697UNR_t$<br>$(-4.36)\quad(-4.34)$ | .30 | 2.04 |
| 3. | $RS_t = -1.375 - .501GAP_t$<br>$(-2.44)\quad(-2.46)$ | .22 | 2.05 |
| 4. | $RS_t = -2.421 + .244GY_t$<br>$(-3.44)\quad(2.22)$ | .20 | 2.11 |
| Sample Period: 1953:I-1978:IV | | | |
| 5. | $RS_t = -3.036 - 3.102E_t$<br>$(-6.89)\quad(-7.54)$ | .47 | 1.97 |
| 6. | $RS_t = -3.978 - 1.692UNR_t$<br>$(-4.22)\quad(-2.74)$ | .35 | 2.17 |
| 7. | $RS_t = -3.152 - .480GAP_t$<br>$(-3.42)\quad(-1.69)$ | .31 | 2.14 |
| 8. | $RS_t = -4.020 + 2.88GY_t$<br>$(-4.56)\quad(2.63)$ | .34 | 2.14 |

[a] See Table 1.

real economic activity. We estimate similar regressions except that we use the same measures of economic activity discussed in the previous section. The results are reported in Table 2 which has the same format as Table 1. In a spirit similar to Fama's, these regression estimates may be interpreted as reflective of the association between real stock returns and real economic activity rather than being indicative of a causal structure. We also adopt the simplification of abstracting from the distinction between anticipated and actual stock returns and real activity, and investigate the relationship between actual variables.[9]

A perusal of the regression results in Table 2 makes it clear that the positive association between stock returns and real activity discovered by Fama is a rather singular outcome which occurs when the growth rate of real *GNP* is the real activity variable. In each of the other cases, the association between real stock returns and economic activity is strongly negative, each of the relevant coefficients being significant at least at the 10 percent level. This finding is true for both time periods, although the results are especially strong for the shorter period. Also, as is the case in regard to Table 1, using the criterion of the size of regression $R^2$, equations with the employment-population ratio variable for economic activity seem superior to others, and these equations appear to present the sharpest contrast to the sign of the correlation between stock returns and real activity postulated by Fama.

Thus, with regard to the relationship between real stock returns and real economic activity, we find, contrary to Fama's hypothesis, considerable empirical evidence that suggests a strong negative relationship. Although not consistent with Fama's position, such a negative association accords well with the conventional view of the relationship.

## III. Discussion

Fama has argued that the observed negative correlation between real stock returns and the rate of inflation is the consequence of a negative association between inflation and real economic activity along with a positive association between real stock returns and real economic activity. He estimated many regression equations in an attempt to empirically support his hypothesis. As he points out, his results are at odds with conventional theories of the Phillips curve and the Mundell-Tobin hypothesis concerning the relationship between these variables. As is well known, Mundell has developed a simple model in which an increase in the expected rate of inflation causes portfolio substitutions from money to financial assets which will reduce the real returns on such assets (for example, stocks). This reduction in real interest rates will stimulate real economic activity. Therefore, according to Mundell's hypothesis, one would expect a positive association between inflation and economic activity and a negative association between real stock returns and economic activity, just the opposite of Fama's conjectures and empirical results but in broad conformity with the empirical results reported in our Tables 1 and 2.

---

[9] See fn. 8. The same remarks broadly apply to both tables.

While these results are consistent with the Phillips curve and the Mundell hypothesis, there remain important questions about the direction of causation. In particular, validity of the Mundell hypothesis would imply unidirectional causation from inflation to real stock returns while Fama's position is that stock returns "lead" the (anticipated) real activity levels which are reflected in (expected) inflation rates. Fama acknowledges the importance of the causality issue, but suggests that his equations, even if true, would probably fail causality tests of the Sims type because one is "not likely to find the exact form of the true structural model or all of the exogenous variables" (1981, p. 556). While there are well-known difficulties associated with these simple tests, it may be instructive to consider what they suggest. Indeed, the causality claims of the Mundell hypothesis are as likely, a priori, to fail such tests.

In this spirit, we have employed Sims' test to explore the causal relationship of fundamental concern: that between inflation and real stock returns. The nature of the test is clearly described by Sims:

> If and only if causality runs one way from current and past values of some list of exogenous variables to a given endogenous variable, then in a regression of the endogenous variable on past, current, and future values of the exogenous variables, the future values of the exogenous variables should have zero coefficients. [p. 541]

We therefore estimated the regression of current real stock returns on current, past, and future rates of inflation and then tested the joint significance of future rates of inflation. We failed to reject the null hypothesis, and therefore the test results lend support to the claim of unidirectional causality from inflation to real stock returns. For further evidence, we also investigated the reverse regression: current inflation on past, future, and current real stock returns. In this case, the hypothesis that the coefficients on future real stock returns are jointly zero was rejected.[10] Together, these results offer significant apparent evidence that, in our sample period, we observe unidirectional causation from inflation to real stock returns. This conclusion is consistent with the Mundell hypothesis, but inconsistent with Fama's postulates and empirical procedures.

## IV. Concluding Remarks

Largely by taking the rate of growth of real *GNP* (or industrial production) as a measure of real economic activity and by deriving the inflation equation from a transformed money demand function, Fama has sought to explain negative correlations between stock returns and inflation as the consequence of a conjunction of a negative association between inflation and real activity and a positive association between stock returns and real activity. In the process, he argued that the Phillips curve hypothesis was not borne out and that the Mundell hypothesis was not justified.

Taking several other plausible variables as measures of real activity and employing an augmented Fisher-Phillips relationship as an inflation equation, we find evidence which suggests that each part of Fama's explanation can be essentially reversed; that there seems, on the whole, evidence of a positive association between inflation and real activity (as theories of the Phillips curve would suggest) and a negative association between real stock returns and real activity (as the Mundell hypothesis would indicate). While we deliberately simplified our procedure by largely abstracting from the distinction between anticipated and actual values, such a simplification does not seem to weaken the argument. The Sims-type "causality" tests conducted by us also tend to support the Mundell hypothesis rather than Fama's. Our explanation is certainly not startling as it is consistent with both the Phillips curve and Mundell-Tobin effects.

---

[10] In performing these tests, variables in the regression equations were first "filtered" in order to produce white noise residuals. Each equation contained four quarter past and future values of the "exogeneous" variable. In the first regression (real stock returns on inflation), the $F$-statistic for the null hypothesis that coefficients on future inflation rates are jointly zero was 1.64. The critical $F(4,75)$ value at the 5 percent significance level is approximately 2.5. Thus we fail to reject the null hypothesis. In the second regression, the $F$-statistic for the null hypothesis that future stock returns have zero coefficients was 3.88, and therefore the hypothesis is easily rejected. Further details are available from the authors.

We do not expect that our less than conclusive evidence will finally settle the issue. Indeed, Fama has raised some questions which have yet to be fully answered. Nevertheless, it would seem premature to abandon prevailing explanations of the negative correlation between the rate of inflation and real stock returns. To the extent that the evidence provided by us is persuasive, it would seem that one need not take recourse to explanations any more tedious than those suggested by Mundell and Tobin.

## APPENDIX: VARIABLES USED AND DATA SOURCES

Inflation ($I_t$): annualized rate (in percentage points) obtained as follows:

$$400\times[ln(CPI)_t - ln(CPI)_{t-3}],$$

where $(CPI)_t$ is the Consumer Price Index ($CPI$-$W$) for January, April, July, and October of each year, and $(CPI)_{t-3}$ is the Consumer Price Index three months before. Data are from U.S. Department of Commerce, *Business Statistics*, 1979.

Nominal rate of interest ($R_t$): the annual rate of return (in percentage points) on Treasury bills three months to maturity on the 15th of January, April, July, and October. Data are from Salomon Brothers. Note that January-April inflation is related to the mid-January estimate of $R_t$, the rate of Treasury bills three months to maturity. Thus our $R_t$ and $I_t$ variables correspond more closley than any other we know of since the closest point of time which the monthly *CPI* may be deemed to reflect is probably the middle of each month.

Employment-population ratio ($E_t$): average for the three months of the calendar quarter as deviation from sample mean. *Source*: Citibank data tape.

Unemployment gap ($UNR_t$): [natural unemployment rate—actual unemployment rate]. *Source*: Robert J. Gordon and updates.

*GNP* gap ($GAP_t$): $100\times$[(Actual real *GNP* − Potential real *GNP*)/Potential real *GNP*]. *Source*: Gordon and updates.

Real *GNP* growth rate ($GY_t$): $(ln Y_t - ln Y_{t-1})\times 400$ where $Y_t$ is real *GNP* for the current quarter. *Source*: *Business Statistics*, 1979.

Real stock returns ($RS_t$): [nominal stock return—rate of inflation]. The nominal stock return data are from the Center for Research in Security Prices of the University of Chicago.

## REFERENCES

**Carlson, John A.**, "Short-Term Interest Rates as Predictors of Inflation: Comment," *American Economic Review*, June 1977, *67*, 469–75.

**Cooley, Thomas F. and LeRoy, Stephen F.**, "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

**Fama, Eugene F.**, "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

______, "Interest Rates and Inflation: The Message in the Entrails," *American Economic Review*, June 1977, *67*, 487–96.

______, "Stock Returns, Real Activity, Inflation, and Money," *American Economic Review*, September 1981, *71*, 545–65.

**Gordon, Robert J.**, *Macroeconomics*, 2d ed., Boston: Little, Brown, 1981.

**Laidler, D. E. W.**, *The Demand for Money: Theories and Evidence*, 2d ed., New York: Harper & Row, 1977.

**Levi, Maurice D. and Makin, John H.**, "Anticipated Inflation and Interest Rates: Further Interpretations of Findings on the Fisher Equation," *American Economic Review*, December 1978, *68*, 801–12.

______ **and** ______, "Fisher, Phillips, Friedman, and the Measured Impact of Inflation on Interest," *Journal of Finance*, March 1979, *34*, 35–52.

**Mehra, Y. P.**, "Is Money Exogenous in Money-Demand Equations?," *Journal of Political Economy*, April 1978, *86*, 211–28.

**Mundell, Robert**, "Inflation and Real Interest," *Journal of Political Economy*, June 1963, *71*, 280–83.

**Plosser, C. I. and Schwert, G. William**, "Estimation of a Noninvertible Moving Average Process: The Case of Overdifferencing," *Journal of Econometrics*, September 1977, *6*, 199–224.

**Sims, Christopher A.**, "Money, Income, and Causality," *American Economic Review*, September 1972, *62*, 540–52.

**Tanzi, Vito**, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates," *American Economic Review*, March 1980, *70*, 12–21.

**Tobin, James**, "Money and Economic Growth," *Econometrica*, October 1965, *33*, 671–84.

**Salomon Brothers**, *An Analytical Record of Yields and Yield Spreads*, New York, 1981.

**U.S. Department of Commerce**, *Business Statistics*, various issues.

# Stock Returns, Real Activity, Inflation, and Money: Reply

*By* EUGENE F. FAMA*

I offer a few remarks on the stock return-inflation relations suggested by Rati Ram and David Spencer, in the hope of encouraging readers to pursue a more detailed comparison of their work and my 1981 article.

1) The Ram-Spencer model is internally inconsistent. They recognize that the model of James Tobin and Robert Mundell that they use implies a negative relation between real activity and the expected real return component of the interest rate.[1] In their equation (5), however, they assume a positive relation between the expected real return and real activity. If they stick with the Tobin-Mundell model, the sign of the real activity coefficient in (6) becomes indeterminate, and their estimates in Table 1 provide no information about my money demand model for inflation.

2) Ram and Spencer give the impression that the only real activity variables used in my tests are industrial production and real *GNP* growth rates. My tables also show results for changes in capital expenditures by nonfinancial corporations and for changes in average real rates of return on their capital stocks. All four real activity measures show strong negative relations with inflation and strong positive relations with real common stock returns. Moreover, stock returns and inflation lead (predict) all of the real activity measures.

3) To justify the alternative inflation model they propose, Ram and Spencer suggest several potential problems in the statistical specification of my money demand model. These specification issues (for example, the role of an interest rate variable) are discussed in my 1981 article, and a reference to the more extensive tests in my 1982 article is provided. None of the concerns raised by Ram and Spencer are evident in the data. However, since they raise the issue, it is interesting to compare the specification of their inflation model with mine.

(a) In the Ram-Spencer inflation model, a Cochrane-Orcutt transformation is needed to obtain well-behaved residuals. In other words, the explanatory variables of their model are not sufficient to explain all of the predictable variation in the inflation rate. In my money demand model for inflation, growth rates of money (the base) and real activity are sufficient to transform the highly autocorrelated inflation rates into white noise residuals.

(b) The coefficients of determination in the Ram-Spencer inflation models include the explanatory power of the lagged inflation rate. Nevertheless, their coefficients of determination are always less than that (.74) from my money demand inflation model, which does not require the lagged inflation rate. My stock return regressions also do not include lagged stock returns as explanatory variables whereas theirs apparently do. The use of lagged returns (via a Cochrane-Orcutt transformation) is curious given the evidence that stock returns show little autocorrelation.

4) The most powerful real activity variable in the Ram-Spencer stock return and inflation regressions is the employment to population ratio. Unlike the real activity growth rates I use, the employment to population ratio is a highly autocorrelated variable that wanders rather smoothly from lower to higher levels during the post-1953 period. (See my 1977 paper.)[2] The ratio is an excel-

---

*Graduate School of Business, University of Chicago.

[1] In the interpretation of their results in Section III, Ram and Spencer state: "As is well known, Mundell has developed a simple model in which an increase in the expected rate of inflation causes portfolio substitutions from money to financial assets which will reduce the real returns on such assets (for example, stocks). This reduction in real interest rates will stimulate real economic activity" (p. 468).

[2] Their other real activity variables, the real *GNP* gap and the difference between the natural and actual rates of unemployment, are likely to manifest similar signs of mean nonstationarity.

lent candidate for a spurious role in regressions where the dependent variables also show evidence of mean nonstationarity. I suspect that this is the role of the ratio in the Ram-Spencer inflation and stock return regressions. In their inflation regressions, which include the level of the nominal interest rate as an explanatory variable, the employment to population ratio probably just proxies (spuriously) for the kind of mean nonstationary behavior of expected real returns documented in my paper with Michael Gibbons, and by Charles Nelson and G. William Schwert (1977), among others. In the stock return regressions, the employment to population ratio probably just proxies for the mean nonstationarity of real stock returns captured in my regressions (but likewise spuriously) by the base growth rate or the expected inflation rate. This is, however, an old issue discussed in my 1977 exchange with John Carlson and Nelson-Schwert in this *Review*, in my 1981 paper, and in my 1982 paper with Gibbons on the Tobin-Mundell model.

## REFERENCES

**Carlson, John A.** "Short-Term Interest Rates as Predictors of Inflation: Comment," *American Economic Review*, June 1977, *67*, 469–75.

**Fama, Eugene F.** "Interest Rates and Inflation: the Message in the Entrails," *American Economic Review*, June 1977, *67*, 487–96.

______, "Stock Returns, Real Activity, Inflation and Money," *American Economic Review*, September 1981, *71*, 545–65.

______, "Inflation, Output, and Money," *Journal of Business*, April 1982, *55*, 201–31.

______ **and Gibbons, Michael R.** "Inflation, Real Returns, and Capital Investment," *Journal of Monetary Economics*, May 1982, *9*, 297–323.

**Mundell, Robert**, "Inflation and Real Interest," *Journal of Political Economy*, June 1963, *71*, 280–83.

**Nelson, Charles R. and Schwert, G. William.**, "Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis that the Real Rate of Interest is Constant," *American Economic Review*, June 1977, *67*, 478–86.

**Ram, Rati and Spencer, David E.**, "Stock Returns, Real Activity, Inflation, and Money: Comment, *American Economic Review*, June 1983, *73*, 463–70.

**Tobin, James**, "Money and Economic Growth," *Econometrica*, October 1965, *33*, 671–84.

# Currency Substitution and Instability in the World Dollar Standard: Comment

*By* MYRON H. ROSS*

In his recent article in this *Review*, Ronald McKinnon states: "*In general, growth in the world money supply is a better predictor of American price inflation than is U.S. money growth*" (p. 324). In support of this, McKinnon gives data in two tables on world money supply percentage increases and rates of inflation in ten industrial countries. He asserts that by lagging his rates of inflation one or two years after the money supply growth rates, this conclusion is supported. Surprisingly, he gives no systematic evidence to support his conclusion.

The evidence, in fact, contradicts McKinnon's conclusion. As shown in Table 1, world money supply is an inferior predictor of U.S. inflation compared with U.S. money supply. For example, with a one-year lag of inflation behind the growth rate of the money supply, U.S. money supply changes and U.S. price level changes have a rank correlation of 0.72 (col. 2). In contrast, world money supply changes and U.S. price level changes have a rank correlation of 0.29. (col. 1) "Domestic monetarism," where the Federal Reserve uses a money growth target (or, for that matter, "domestic Keynesianism" which uses an interest rate target), may be inefficient as McKinnon argues, because it ignores the foreign exchanges. However, the evidence appears to indicate that ignoring the foreign exchanges is probably not a significant shortcoming of "domestic monetarism."

TABLE 1—SPEARMAN RANK CORRELATION COEFFICIENTS OF PERCENTAGE CHANGES IN MONEY SUPPLY AND PERCENTAGE CHANGES IN THE PRICE LEVEL, 1960–80[a]

| Money:<br>Price Level: | World<br>U.S. | U.S.<br>U.S. | World<br>World |
|---|---|---|---|
| One-Year Lag | 0.29 | 0.72 | 0.35 |
| Two-Year Lag | 0.59 | 0.76 | 0.58 |

[a]All data are from McKinnon's Tables 1 and 2. The money supply is $M1$, and the price level is the wholesale price index. The ten industrial countries are: United States, Canada, Japan, United Kingdom, Germany, France, Italy, Netherlands, Belgium, and Switzerland. To determine mean money supply growth rates and mean rates of inflation, each country has a weight equal to 1970 nominal *GNP*.

## REFERENCE

**McKinnon, Ronald I.,** "Currency Substitution and Instability in the World Dollar Market," *American Economic Review*, June 1982, *72*, 320–33.

---

*Western Michigan University.

# Currency Substitution and Instability in the World Dollar Standard: Reply

*By* RONALD MCKINNON AND KONG-YAM TAN*

Myron Ross correctly points out in his comment that U.S. price inflation is better predicted by the American money supply ($M1^{US}$) than by the broader ten-country world money supply ($M1^{W}$) over the whole statistical time-series from 1960 to 1980 provided in McKinnon's 1982 article. Specifically, he showed that American price inflation is more highly correlated with $M1^{US}$ lagged one or two years than with $M1^{W}$ similarly lagged.

However, McKinnon's present-tense assertion that "In general, growth in the world money supply is a better predictor of American price inflation than is American money growth" applies only to the weak dollar standard of the 1970's and early 1980's—as his preceding discussion intended, but failed to indicate clearly. Only in this later period of volatile exchange rates and price-level instability in the United States do alternative hard currencies (such as the yen and deutsche mark) become competitive as international stores of value and units of account. Hence, international currency substitution—associated with the ebb and flow of speculation against or for the dollar—significantly destabilized the demand for $M1^{US}$ in the 1970's and 1980's. And only in this later period might one expect the sum of these internationally substitutable monies, $M1^{W}$, to predict world, and perhaps even American, price inflation better than does $M1^{US}$.

In contrast, during the strong dollar standard of the 1950's and 1960's, the dollar was unchallenged as international money. Exchange rates were (by and large) convincingly fixed: speculation for or against the dollar was incapable of substantially altering *U.S.* interest rates or of directly affecting the demand for $M1^{US}$. Because cyclical international influences did not then destabilize the demand for dollars, $M1^{US}$ was by itself a fairly efficient predictor of American prices. Thus Ross's statistical results for the period 1960 to 1980 show somewhat greater predictive strength in $M1^{US}$, compared to $M1^{W}$, because the 1960's data outweigh the dissimilar data of the 1970's.

Let us instead partition the sample of IMF data around the year 1970, whence began the transition from fixed to fluctuating exchange rates and the maturation of alternative monetary systems in Europe and Japan. After applying similar statistical correlation procedures to those used by Ross, we show a striking structural shift in the world's monetary system: the explanatory power of $M1^{W}$ increases sharply as that of $M1^{US}$ declines moderately (Tan, 1982).

For the early period of 1960 to 1970, Table 1 shows simple correlation coefficients between annual percentage changes in money supplies and changes in American and world wholesale price indices. The $M1^{US}$ provides a good explanation of U.S. prices *and* of world prices one to two years hence, whereas the broader definition of $M1^{W}$ (in which $M1^{US}$ enters with a 50 percent weight) does surprisingly poorly, being insignificantly correlated with American or world prices. Apparently, money growth in industrial countries other than the United States was not then an independent source of worldwide inflationary pressure during the strong dollar standard.

Table 2 shows the same simple correlations between percentage changes in prices and money for the weak dollar standard of

---

*Stanford University. Please note that a typographical error giving the last word in the original title as "Market" appeared on the cover of the June 1982 issue of this *Review*. The title was correct in the table of contents and on the article.

TABLE 1—SIMPLE CORRELATIONS BETWEEN WHOLESALE PRICES AND LAGGED MONEY SUPPLIES: 1960–70

| Money Supply:<br>Price Level: | U.S.<br>U.S. | U.S.<br>World | World<br>U.S. | World<br>World |
|---|---|---|---|---|
| One-Year Lag | 0.71 | 0.52 | 0.13 | 0.03 |
| Two-Year Lag | 0.64 | 0.76 | −0.16 | 0.04 |

*Source:* Percentage charges in money supplies and in prices taken from Tables 1 and 2 of McKinnon.

TABLE 2—SIMPLE CORRELATIONS BETWEEN WHOLESALE PRICES AND LAGGED MONEY SUPPLIES: 1970–81

| Money Supply:<br>Price Level: | U.S.<br>U.S. | U.S.<br>World | World<br>U.S. | World<br>World |
|---|---|---|---|---|
| One-Year Lag | 0.46 | 0.35 | 0.12 | 0.05 |
| Two-Year Lag | 0.71 | 0.67 | 0.86 | 0.86 |

*Source:* Same as Table 1. Data for 1981 calculated similarly from IMF *International Financial Statistics.*

1970–81.[1] The explanatory power of $M1^{US}$ declines somewhat—rather more in its ability to predict world prices in comparison to the earlier period. More striking is the increased ability of $M1^{W}$, lagged two years to predict both world and American prices: in Table 2 the simple correlation coefficients are .86 for both—which is higher than the equivalent coefficients for $M1^{US}$. On the other hand, $M1^{W}$ appears to have little or no predictive power after a one-year lag. Because the monetary data are end-of-year stocks while the price data are over-the-year averages, what appears as a two-year lag in the tables is really closer to one and one-half years; and what appears as a one-year lag is effectively only six months.

After this transition from a strong to a weak dollar standard, these statistics suggest that the money supplies of other industrial countries, which make up approximately one-half of the statistical index $M1^{W}$, now exert a stronger independent influence on both world and American prices. Whether or not $M1^{W}$, when used by itself, predicts American prices better than does $M1^{US}$ is not critical to establishing that such an important structural change has actually occurred. Nevertheless, McKinnon's earlier assertion quoted by Ross, that $M1^{W}$ by itself is better than $M1^{US}$ by itself for predicting American prices, is valid for the period 1970–81 of the weak dollar standard.

Testing this assertion requires the use of multiple regressions that simultaneously allow for both one- and two-period lags in the impact of monetary changes. Table 3 shows

TABLE 3—REGRESSION OF U.S.. WHOLESALE PRICES ON AMERICAN AND WORLD MONEY STOCKS (LAGGED ONE AND TWO YEARS)

| | Constant | $M1^{US}_{-1}$ | $M1^{US}_{-2}$ | $\bar{R}^2$ | D-W | Period |
|---|---|---|---|---|---|---|
| $P^{US}$ | −1.76 | 0.46 | 0.39 | 0.59 | 1.56 | 1960–70 |
| | (−1.78) | (2.44) | (2.42) | | | |
| | −5.86 | 0.82 | 1.54 | 0.53 | 1.02 | 1970–81 |
| | (−1.41) | (1.58) | (3.06) | | | |
| | | $M1^{W}_{-1}$ | $M1^{W}_{-2}$ | | | |
| | 2.71 | 0.03 | −0.16 | −0.25 | 0.85 | 1960–70 |
| | (0.42) | (0.06) | (−0.31) | | | |
| | −6.52 | −0.13 | 1.88 | 0.70 | 1.86 | 1970–81 |
| | (−1.66) | (−0.40) | (5.04) | | | |

*Source:* Same as Table 2. *Note: t*-statistics are shown in parentheses. Data on the independent variable, $P^{US}$, start two years into each period.

---

[1] In McKinnon's earlier article, IMF data up to the end of 1980 were used. Adding more recent data for 1981 does not significantly change any of the correlation coefficients.

the consequences of regressing changes in *U.S.* prices on $M1^{US}_{-1}$ and $M1^{US}_{-2}$, and then on $M1^{W}_{-1}$ and $M1^{W}_{-2}$. For 1970–81, the $\bar{R}^2$ adjusted for degrees of freedom in the multiple regression equation using $M1^{US}$ is 0.53 and the $\bar{R}^2$ using $M1^{W}$ is 0.70. To check that this difference is significant, we did an $F$ test by comparing the total explained variance in American prices when only $M1^{US}$ was used (lagged one and two periods) to that explained by an equation using both $M1^{US}$ (lagged one and two periods) and the incremental variable $M1^{W}-M1^{US}$ (lagged two periods). The money supplies of other industrial countries were thus found, at the 95 percent level of confidence, to add significantly to the explanation of American prices.

Other information from the foreign exchanges remains to be incorporated. Nevertheless, that contained in $M1^{W}$ alone suggests that the doctrine of domestic monetarism—where American authorities consider only American monetary aggregates—has become obsolete.

## REFERENCES

**McKinnon, Ronald I.**, "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, *72*, 320–33.

**Ross, Myron J.**, "Currency Substitution and Instability in the World Dollar Standard: Comment," *American Economic Review*, June 1983, *73*, 473.

**Tan, Kong-Yam**, "World Money, Income and Prices: Some Preliminary Evidence," unpublished, Stanford University, April 1982.

# The Short-Run Relation Between Growth and Inflation in Latin America: Comment

*By* SEBASTIAN EDWARDS*

In analyzing the relationship between unanticipated monetary policy and growth for a group of Latin American countries, using the approach of Robert Lucas (1973) and Robert Barro (1977), James Hanson concludes that for these countries, "[A]s a rule of thumb, ten percentage points of unexpected inflation raise output about one percentage point above trend..." (1980, p. 987). However, Hanson's results are not robust and depend critically on the functions he uses to generate expectations of monetary growth. If more general monetary processes are used, and open economy considerations are introduced into the analysis, Hanson's results are substantially changed.

The purpose of this comment is to show that when a monetary process that explicitly incorporates the role of fiscal deficits is considered, Hanson's general result—suggesting an elasticity of .10 between output growth and unexpected money for all these countries—does not hold.[1] In particular, this alternative monetary process indicates that these countries exhibit very different behavior with respect to the relationship between unexpected money and growth. Furthermore, Lucas' proposition of an inverse relationship between the effect of (unexpected) monetary policy and the variability of the money supply seems to hold for these countries. My paper also incorporates into the analysis the fact that these countries are open economies. This is done in three ways: 1) changes in the terms of trade are explicitly incorporated as possible determinants of growth; 2) for the case of Mexico—which maintained a fixed exchange rate during most of the period—domestic credit is also used as the relevant policy variable; 3) the possibility that all these countries are subject to *common* external shocks is incorporated explicitly in the estimation procedure.[2]

Hanson's study is based on the estimation of equations for Brazil, Chile, Colombia, Mexico, and Peru of the following type:

$$(1) \qquad DY_t = a_2 + a_1 DMR_t + u_t,$$

where $Y_t$ is the *log* of real output, $DY_t = (log\, Y_t - log\, Y_{t-1})$, and $DMR_t$ is the unexpected change in the *log* of the quantity of money. Hanson tries several specifications for $DMR_t$. Some of his best results are obtained when $DMR_t$ is replaced by actual changes in the quantity of money $DM_t$ (see his Table 1), and when $DMR_t$ is defined as actual growth of the quantity of money minus past inflation—$DMR_t = DM_t - DP_{t-1}$ (see his Table 4). Table 1 presents the results obtained from the estimation of equation (1) using these definitions of $DMR_t$. Even though the data used do not correspond exactly to

---

*University of California-Los Angeles, and National Bureau of Economic Research. The research reported here is part of the NBER's research program in International Studies. Opinions expressed are my own and not those of the Bureau. I am indebted to Michael Darby, Alejandra Cox, Max Fry, and an anonymous referee for helpful comments.

[1] On the relationship between money creation and fiscal deficits see, for example, Arnold Harberger (1964, 1978), Ricardo Ffrench-Davis (1973), Bijan Agheveli and Mohsin Khan (1978) and Werner Baer and Paul Beckerman (1974). Hanson (1980, p. 975) acknowledges the importance of money creation as a source of government revenue in Latin America. However, he does not incorporate the fiscal deficit as an explanatory variable in his money supply processes.

[2] On the relationship between growth and monetary policy in an open economy using a Lucas-Barro type of framework see Leonardo Leiderman (1979), Mario Blejer and Roque Fernandez (1980) and Michael Darby (1983). Barro (1978) also included a terms of trade variable in his study of the U.S. case. An alternative way of introducing open economy considerations is to include the growth of international reserves in the money supply equation. Results obtained using this formulation for the case of Latin America do not alter the results presented in this paper.

TABLE 1—GROWTH OF OUTPUT AND MONEY GROWTH IN LATIN AMERICA

| | Brazil 1952–74 | Chile 1952–70 | Colombia 1951–74 | Mexico 1950–74 | Peru 1951–74 |
|---|---|---|---|---|---|
| A. $DY_t = a_2' + a_1 DM_t + w_t$ | | | | | |
| $a_2'$ | .107 | .040 | .018 | .036 | .033 |
| | (4.340) | (1.930) | (1.107) | (2.889) | (3.019) |
| $a_1$ | −.103 | −.012 | .190 | .227 | .131 |
| | (1.481) | (.198) | (1.926) | (2.220) | (2.448) |
| $R^2$ | .109 | .003 | 0.171 | .206 | .261 |
| $D\text{-}W$ | 1.123 | 1.303 | 2.047 | 2.046 | 1.227 |
| B. $DY_t = a_2^{*\prime} + a_1^*(DM_t - DP_{t-1}) + w_t^*$ | | | | | |
| $a_2^{*\prime}$ | .063 | .032 | .049 | .045 | .042 |
| | (7.428) | (6.258) | (8.821) | (8.197) | (5.906) |
| $a_1^*$ | .159 | .142 | .002 | .239 | .148 |
| | (2.073) | (3.035) | (.031) | (3.711) | (2.748) |
| $R^2$ | .193 | .365 | .001 | .420 | .308 |
| $D\text{-}W$ | 1.400 | 2.263 | 1.825 | 2.074 | 1.192 |

*Note:* The data sources are described in the Appendix. Absolute $t$-statistics are shown in parentheses. $R^2$ refers to the coefficient of correlation and $D\text{-}W$ is the Durbin-Watson statistic.

Hanson's, the results are very similar.[3] These results were obtained, as in Hanson's paper, using a narrow definition of money ($M_1$). When a broad definition of money is used, however, the results are similar. These results, together with pooled estimations, constitute the base of Hanson's analysis. However, as is shown below, once fiscal deficits and open economy factors are introduced, the similarity across countries of output response to monetary changes reported in Table 1 tends to disappear.

## I. Fiscal Deficit and Monetary Growth in Latin America

A well-known feature of developing countries in general, and Latin American countries in particular, is that money creation is an important source of government revenue (see, for example, Harberger, 1964, 1978; Ffrench-Davis; and Baer-Beckerman). In this section, I present results obtained from the estimation of monetary growth equations that incorporate explicitly the role of the fiscal deficit in money creation. Specifically, it is assumed that in every period the growth of the quantity of money responds partially to the current fiscal deficit, and to past rates of money creation:

$$(2)\quad DM_t = \alpha_0 + \sum_{i=1}^{k} \alpha_i DM_{t-i} + \beta DEF_t + w_t,$$

where $DEF_t$ is the ratio of the fiscal deficit to the quantity of money in $t-1$. The residuals obtained from the estimation of this equation [$\hat{w}_t = DM_t - \widehat{DM}_t$] are used in Section II as a measure of unexpected money.[4]

---

[3] I have followed, as far as possible, the indications in Hanson's (p. 979) paper to construct my data. However, I have deliberately introduced some changes. In particular, I have used *ECLA*'s data on *GDP* for Brazil and Mexico, since it is well known that the *IFS* data set contains serious flaws in these series for the earlier periods. In private communication, Hanson has indicated to me that his monetary data for Colombia was taken from the Banco de la Republica Bulletin. In this comment, however, the monetary data for Colombia and other countries—except Chile—are yearly averages constructed from the *IFS* data. See the Appendix for further details on the data. The nonsignificance of $a_1^*$ for Colombia contrasts with Hanson's results. The reason for this is the different monetary series used.

[4] It is important to note that the use of equations of this type to generate expectations of money growth implicitly assumes that their parameters are stable through the period. The reason for this is that, as Barro (1977, p. 105–06) has pointed out, this procedure uses future information (not available at $t-i$) to generate the estimated $\hat{\alpha}$ parameters. If these parameters are not stable, however, a period-by-period sequential updating procedure should be implemented.

TABLE 2—MONETARY GROWTH AND FISCAL DEFICIT IN LATIN AMERICA

$$DM_t = \alpha_0 + \sum_{i=1}^{4} \alpha_i DM_{t-i} + \beta DEF_t + u_t$$

| Equation | | Monetary Aggregate | Constant | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\beta$ | $SEE$ | $D\text{-}W$ | $R^2$ | $F$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (2.1) | Brazil | $M2$ | .054 | .904 | −.529 | .343 | — | .392 | .0530 | 1.858 | .829 | 18.17 |
| | | | (1.455) | (4.585) | (1.992) | (1.870) | | (3.285) | | | | |
| (2.2) | Chile | $M2$ | .138 | .739 | −.358 | — | — | .083 | .0797 | 1.632 | .427 | 3.98 |
| | | | (.880) | (2.844) | (1.175) | | | (.709) | | | | |
| (2.3) | Colombia | $M2$ | −.064 | .598 | −.067 | .730 | — | .576 | .0336 | 1.918 | .594 | 5.60 |
| | | | (.878) | (3.152) | (.279) | (2.538) | | (3.719) | | | | |
| (2.4) | Mexico | $M1$ | .107 | .901 | −.971 | .212 | — | .088 | .0189 | 1.676 | .746 | 11.75 |
| | | | (2.993) | (4.507) | (5.406) | (1.262) | | (2.487) | | | | |
| (2.5) | Mexico | Domestic Credit | .021 | .987 | −.806 | .696 | — | −.403 | .0592 | 1.807 | .525 | 10.87 |
| | | | (.307) | (3.563) | (3.042) | (2.129) | | (.424) | | | | |
| (2.6) | Peru | $M1$ | .085 | .164 | −.329 | .062 | .332 | .343 | .0782 | 1.631 | .613 | 3.80 |
| | | | (1.353) | (.790) | (1.519) | (.291) | (1.573) | (3.165) | | | | |

*Notes:* Absolute $t$-statistics in parentheses. *SEE* refers to standard error of the regression.

Table 2 presents the estimates of equation (2) for five Latin American countries. In most cases, $k$ equals 3. For Chile, however, $k$ equals 2, and for Peru, $k$ equals 4, to produce white noise residuals. In addition, for Mexico—the only country in the sample that had a fixed exchange rate for most of the period—an equation for domestic credit is also reported.[5] As may be seen, in most cases the estimated coefficient for $DEF_t$ is positive and significant. Also note that in all cases, the $F$-statistic indicates that the regression is significant at conventional levels. The Durbin-Watson statistic, and analysis of the autocorrelation functions of the residuals using the Box-Jenkins procedure show that the residuals of these monetary equations are white noise, indicating that they are appropriate candidates for the measure of unexpected money. An important characteristic of the money equations presented in Table 2 is that, when measured by the standard deviation of the prediction error, they outperform simple autoregressions and past inflation in predicting actual changes in the money stock.[6]

## II. Monetary Shocks and Growth in Latin America

This section reexamines the relationship between growth and unexpected monetary policy in Latin America, using the residuals from the money supply equations reported in Table 2 as a measure of unexpected money. The possible role of changes in terms of trade ($DTOT_t$) in these countries' growth is also investigated. For each country, the following equation was estimated:

$$(3) \quad DY_t = \delta_0 + \sum_{i=0}^{k} \gamma_i DMR_{t-i} + \theta DTOT_t + (\phi\ TIME) + \varepsilon_t,$$

where $DMR_{t-i}$ are residuals from the estimation of equation (2).[7] The time variable is

---

[5] For all countries, equations for $M_1$, $M_2$, and domestic credit were fitted. However, due to space considerations, only the "best" results—as measured by the standard error of the regression—are reported in Table 2. For the cases of Brazil, Colombia, and Peru, the residuals from the domestic credit equations exhibited serially correlated errors.

[6] With the exception of one case, Chile (when three-period lagged inflation is used to predict money), these equations' predictive power outperforms those used by Hanson.

[7] It is important to note that the parameters of equations (2) and (3) are related, and that both could be estimated simultaneously imposing cross-equation restrictions. Results obtained using this approach for these Latin American countries confirm the main conclusions presented herein.

TABLE 3—OUTPUT GROWTH AND UNEXPECTED MONEY IN LATIN AMERICA: ORDINARY LEAST SQUARES RESULTS
$DY_t = \delta_0 + \Sigma\gamma_i DMR_{t-i} + \theta DTOT_t + \phi TIME + \varepsilon_t$

| Equation | | Monetary Aggregate | $\delta_0$ | $\gamma_0$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\theta$ | $\phi$ | $D\text{-}W$ | $R^2$ | $F^a$ | $F^b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (3.1) | Brazil | $M2$ | −6.106 | −.040 | −.087 | −.423 | −.273 | .023 | .003 | 1.813 | .546 | 1.60 | 1.02 |
| | | | (1.235) | (.176) | (.382) | (1.775) | (1.252) | (1.136) | (1.248) | | | | |
| (3.2) | Chile | $M2$ | .035 | −.031 | −.064 | −.086 | −.012 | .163 | — | 1.587 | .413 | .99 | .35 |
| | | | (4.609) | (.348) | (.581) | (.866) | (.132) | (2.075) | | | | | |
| (3.3) | Colombia | $M2$ | .055 | .153 | .113 | .021 | .289 | .028 | — | 2.367 | .419 | 1.30 | 1.11 |
| | | | (13.474) | (1.093) | (.669) | (.116) | (1.566) | (.611) | | | | | |
| (3.4) | Mexico | $M1$ | .073 | .472 | .741 | .526 | .176 | .188 | — | 1.966 | .484 | 1.69 | 1.35 |
| | | | (11.835) | (1.494) | (2.139) | (1.326) | (.472) | (1.961) | | | | | |
| (3.5) | Mexico | Domestic Credit | .063 | −.084 | −.060 | −.222 | .024 | .116 | — | 1.998 | .377 | 1.09 | .73 |
| | | | (11.114) | (.680) | (.482) | (1.667) | (.207) | (1.164) | | | | | |
| (3.6) | Peru | $M1$ | .053 | .303 | .118 | .054 | .064 | .062 | — | 1.861 | .524 | 1.98 | 2.23 |
| | | | (8.269) | (2.750) | (1.253) | (.600) | (.689) | (.949) | | | | | |

*Notes:* Absolute $t$-statistics in parentheses.

[a] $F$-test for the significance of the regression.

[b] $F$-test for the significance of the $\gamma_i$s as a group. The null hypothesis is $\gamma_i = 0$, $i = 0, \ldots, 3$.

included only for Brazil in order to reduce the degree of serial correlation observed in the residuals.

In the analysis, $k$ ranges from zero to three. There are several ways to justify the inclusion of lagged values of $DMR$ in equation (3). First, as Barro (1978, p. 553) has argued, lagged $DMR$s may capture the effects of unexpected monetary shocks on stock variables (like capital) that are carried forward. The inclusion of lagged values of $DMR$ also pick up both short- and long-run effects of monetary shocks on growth. While $\gamma_0$ captures the short-run effect of unexpected monetary shocks, $\Sigma_{i=0}^{k}\gamma_i$ captures the long-run effect of these shocks. This interpretation of the coefficients of lagged values of $DMR$ in an equation of the type of (3) has recently been suggested by Roger Kormendi and Philip Meguire (1981). Furthermore, if the natural rate hypothesis is true, $\Sigma_{i=0}^{k}\gamma_i$ should not be significantly different from zero.

Equation (3) was estimated in two alternative ways. First, $OLS$ were applied for each country, using the residuals from the monetary process equations reported in Table 2 as measures of unexpected money. Second, equation (3) was also estimated using Arnold Zellner's (1962) seemingly unrelated regression procedure ($SURE$) for the cases of Brazil, Colombia, Mexico, and Peru for 1954–74. Since all these countries are open economies, and hence subject to common external shocks, the residuals ($\varepsilon_t : t = 1, T$) from the growth equations for each of them could be correlated. The use of this $GLS$ procedure is superior to simply pooling the different equations since it still allows for differences in the $\gamma_i$s across countries.

The results obtained from the $OLS$ estimate of equation (3) are reported in Table 3. Table 4, on the other hand, contains the seemingly unrelated regression results from the simultaneous estimation of equations (3.1), (3.3), (3.4), and (3.6) for Brazil, Colombia, Mexico, and Peru.[8] As can be seen, the results are very different across countries. For Brazil and Chile, no evidence was found of a significant effect of unexpected money on output growth. On the other hand, some evidence of significant positive effects of unexpected money on growth was found for Colombia, Mexico, and Peru. The significant coefficients for unexpected money changes,

---

[8] Chile was not included in the $SURE$ estimations in order for the estimations to be based on a longer time period. Other sets of equations were also simultaneously estimated using $SURE$. The results are not reported here due to space considerations. Equation (3) was also estimated in terms of levels. The results obtained support the conclusions presented in this paper.

TABLE 4—OUTPUT GROWTH AND UNEXPECTED MONEY IN LATIN AMERICA: SEEMINGLY UNRELATED REGRESSIONS RESULTS

$$DY_t = \delta_0 + \Sigma\gamma_i DMR_{t-i} + \theta DTOT_t + \phi TIME + \varepsilon_t$$

| Equation | | Monetary Aggregate | $\delta_0$ | $\gamma_0$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\theta$ | $\phi$ |
|---|---|---|---|---|---|---|---|---|---|
| (4.1) | Brazil | $M2$ | −6.134 | −.048 | −.052 | −.445 | −.296 | .020 | .003 |
| | | | (1.312) | (.217) | (.235) | (1.940) | (1.409) | (1.001) | (1.326) |
| (4.3) | Colombia | $M2$ | .055 | .195 | .069 | .060 | .298 | .016 | — |
| | | | (13.718) | (1.839) | (.481) | (.404) | (2.058) | (.412) | |
| (4.4) | Mexico | $M1$ | .074 | .246 | .632 | .538 | .202 | .269 | — |
| | | | (12.749) | (.942) | (2.237) | (1.553) | (.647) | (3.637) | |
| (4.6) | Peru | $M1$ | .052 | .325 | .138 | .066 | .051 | .060 | |
| | | | (8.198) | (2.965) | (1.481) | (.751) | (.553) | (.941) | |

*Note:* Absolute *t*-statistics in parentheses.

however, vary widely in both magnitude and lags.[9] The results for Mexico are particularly interesting. While none of the $\gamma_i$s is significant when unexpected changes in domestic credit are used, a positive, significant coefficient for $\gamma_1$ is found when $M_1$ is used as the relevant monetary aggregate. This is, to some extent, surprising, since one would expect that in the case of a small open economy with fixed exchange rates, domestic credit would be the relevant policy variable.[10] The results presented in Tables 3 and 4 also indicate that changes in the terms of trade are positively related to output growth in these countries ($\theta$ is positive in all cases). These coefficients, however, are only significant at the conventional levels for the cases of Chile and Mexico.

## III. Money, Inflation, and Growth in Brazil and Chile

According to the discussion presented in the preceding section for Brazil and Chile, unexpected monetary changes have no effects on output growth. These results, however, appear to be inconsistent with Hanson's findings—reported in panel B of my Table 1. This inconsistency is more apparent than real, and is due to the peculiar definition of unexpected money used by Hanson.[11] In order to investigate this problem further, I reestimated Hanson's equations without restricting the coefficients of $DM_t$ and $DP_{t-1}$ to be equal and opposite. The results for Brazil and Chile are

Brazil:

$$DY_t = \underset{(4.439)}{.098} + \underset{(.536)}{.050}\ DM_t - \underset{(2.250)}{.159}\ DP_{t-1},$$

$$D\text{-}W = 1.747;\ R^2 = .313;\ F = 3.87.$$

Chile

$$DY_t = \underset{(1.980)}{.040} + \underset{(.611)}{.045}\ DM_t - \underset{(1.361)}{.065}\ DP_{t-1},$$

$$D\text{-}W = 2.167;\ R^2 = .119;\ F = .95.$$

In both cases, the coefficient of actual money changes is not significant, while the coefficient for past inflation is significantly *negative* for Brazil. Furthermore, the restriction imposed by Hanson of equality (with opposite signs) of the coefficients of $DM_t$ and $DP_{t-1}$ is rejected at the conventional levels for Brazil. When these results are looked at from this perspective, they are perfectly consistent with my above findings in Section II: in these highly inflationary countries, money growth (actual and unexpected) has *no* effect on output growth. However, inflation tends to have a significantly *negative* effect on growth in these countries.

---

[9] As may be seen in Table 3, the null hypothesis that the $\gamma_i$s are significant as a group cannot be rejected at the conventional levels. Also, for all countries except Peru, the hypothesis $\Sigma_{i=0}^{k}\gamma_i = 0$ cannot be rejected at the 5 percent level.

[10] Blejer and Fernandez, in their study on Mexico that distinguishes between tradable and nontradable goods, have found a positive effect of unexpected increases in domestic credit on nontradables output.

[11] Note, however, that theoretically, under certain restrictive assumptions (expected money growth equals growth in nominal money demand, and expected rate of inflation equals past inflation), it is possible to relate $(DM_t - DP_{t-1})$ to unexpected money growth.

## APPENDIX

*Real Income*: for Colombia it is real *GDP* taken from the *International Financial Statistics* (*IFS*); for Peru and Brazil, real *GDP* obtained from U.N. *National Accounts* was used; for Mexico, real *GDP* from *ECLA*, as reported in UCLA's *Statistical Abstract of Latin América*, was used; for Chile, data on real *GDP* taken from Ffrench-Davis were used.

*Money* (*M1 and M2*) *and Domestic Credit*: for Brazil, Colombia, Mexico, and Peru, yearly averages constructed from the *IFS* raw data were used. For Chile, the series reported in Ffrench-Davis was used.

*Prices*: for Brazil, Colombia, Mexico, and Peru, the data were taken from the *IFS*; for Chile, the Ffrench-Davis series was used.

*Terms of Trade*: taken from *ECLA*'s "América Latina: Relación de Términos de Intercambio, 1928–1976."

*Fiscal Deficit*: for Brazil, Colombia, and Peru, the data were taken from the *IFS*; for Chile, data from Ffrench-Davis were used; for Mexico, *Anuario Estadístico Compendiado* (various issues).

## REFERENCES

**Agheveli, Bijan B. and Khan, Mohsin S.**, "Government Deficit and Inflationary Process in Developing Countries," *IMF Staff Papers*, September 1978, *25*, 383–416.

**Baer, Werner W., and Beckerman, Paul**, "Indexing in Brazil," *World Development*, October/December 1974, *2*, 35–48.

**Barro, Robert**, "Unanticipated Money Growth and Employment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

______, "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August 1978, *86*, 549–80.

**Blejer, Mario J. and Fernandez, Roque B.**, "The Effects of Unexpected Money Growth on Prices and on Output and its Composition in a Fixed Exchange Rate Open Economy," *Canadian Journal of Economics*, February 1980, *13*, 82–95.

**Box, George E. P., and Jenkins, Gwilym M.**, *Time Series Analysis*: *Forecasting and Control*, 2d ed., San Francisco: Holden-Day, 1976.

**Darby, Michael R.**, "Unanticipated or Actual Changes in Aggregate Demand Variables: A Cross-Country Analysis," in his et al., *The International Transmission of Inflation*, forthcoming 1983.

**Ffrench-Davis, Ricardo**, *Políticas Económicas en Chile, 1952–70*, Santiago: Editorial Nueva Universidad, 1973.

**Hanson, James**, "The Short-Run Relation Between Growth and Inflation in Latin America," *American Economic Review*, December 1980, *70*, 972–89.

**Harberger, Arnold C.**, "Notes on Inflation," in W. Baer and I. Kenstenetzky, eds., *Inflation and Growth in Latin America*, Illinois: Homewood, 1964.

______, "A Primer on Inflation," *Journal of Money, Credit and Banking*, August 1978, *10*, 505–21.

**Kormendi, Roger C. and Meguire, Philip G.**, "The Real Output Effect of Monetary Shocks: Cross Country Tests of Rational Expectations Propositions," Report 8123, Center for Mathematical Studies in Business and Economics, University of Chicago, 1981.

**Leiderman, Leonardo L.**, "Expectations and Output-Inflation Tradeoffs in a Fixed-Exchange-Rate Economy," *Journal of Political Economy*, December 1979, *87*, 1285–306.

**Lucas, Robert E., Jr.**, "Some International Evidence on the Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

**Zellner, Arnold G.**, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, June 1962, *57*, 348–68.

# The Short-Run Relation Between Growth and Inflation in Latin America: Reply

By JAMES A. HANSON*

Sebastian Edwards uses a new, "unexpected" money variable—based on the addition of the contemporaneous fiscal deficit to my autoregressive money growth equations—to argue that 1) in Brazil and Chile there is no significant relationship between unexpected money growth and output growth, 2) unexpected money growth has different estimated effects on output in Colombia, Peru, and Mexico, and these effects substantially exceed those estimated in Brazil and Chile, and 3) thus my 10 percent rule for the effect of unexpected money growth on output growth does not hold. In addition, Edwards incorporates the terms of trade in his output equations and finds they are important in the case of Chile and Mexico.[1] Finally, Edwards takes issue with my use of lagged inflation to explain output.

Edwards and I use somewhat different data, not only in terms of sources, but also Edwards often uses $M2$ rather than $M1$ in reporting his best results. However, Edwards' and my basic results are broadly similar (compare his Table 1 and my original Tables 1 and 4). This is encouraging, and in general makes it seem not worth quibbling over the differences in data. However, there is one difference which affects the results that I will discuss later.

The inclusion of the fiscal deficit in the money growth equation is a key difference in our work. As I pointed out in my earlier article, fiscal and monetary policy often are closely related in Latin America. Thus it is not surprising that Edwards finds the deficit (divided by lagged money) is a significant factor in explaining money growth. However, the point is *not* to explain money growth, but to estimate a relationship which the public could have been using to predict how much money, and thus nominal aggregate demand, has grown since the last observation on the money stock.[2]

It strains credulity to assume that, in Latin America, the budget deficit could be used by the public to estimate money growth contemporaneously. Data on the cash deficit usually are available publicly only with a long lag. Moreover, many alternative versions of a deficit, involving creative accounting methods, are reported. For example, for many years in Colombia, the reported deficit, which is available with about a one-quarter lag, treated both the unspent portion of the budget and the accounting profits from the revaluation of the local currency value of exchange reserves as income; procedures which caused great confusion inside as well as outside the government. Despite such problems, the cash deficit still might be useful as a monetary growth predictor; for example, it may be correlated with some other easily available contemporaneous indicator of money growth. However, its usefulness can best be judged from a comparison of the results when the deficit is used only as the money growth predictor to determine unexpected money (a constrained equation) with those obtained when it is used as an additional independent variable (an unconstrained equation), a test Robert Barro has suggested in another context.

---

*World Bank. The views and interpretations expressed in this article are my own and should not be attributed to the World Bank, its affiliated organizations, or to any individual acting in their behalf.

[1] Edwards also uses the seemingly unrelated regression technique to take into account possible interrelationships in the error terms across countries. Judging from the reported *t*-statistics, this procedure did not improve the fits greatly.

[2] Surprisingly, the deficit is not significant in explaining domestic credit in Mexico. This result and the estimated negative relationship between Edwards' domestic credit series and output suggest a problem in the domestic credit variable.

Lacking such a comparison, it is still possible to make some judgement by comparing Edwards' output results, which include the fiscal deficit as a monetary predictor, with my results excluding it. Aside from some rather large coefficients for Mexico, Edwards' results (Table 3) are broadly similar to those in my Table 2—unexpected monetary growth has a negative, insignificant impact on output growth in Brazil and Chile, and positive effects in Colombia and Peru. The differences are only that 1) Edwards includes some lagged unexpected money terms, 2) in the cases of Colombia, Peru and, especially Mexico, his coefficients are somewhat larger than I obtained, and 3) the time pattern of significance differs, though given the significance of his coefficients it may be hard to reject the hypothesis of only contemporaneous effects of unexpected money growth. Why then do we obtain such different conclusions?

I did *not* claim that the countries responded similarly to unexpected money growth when autoregressive predictors of money growth were used in all five countries. Indeed, such a result would refute the Lucas hypothesis since both my and Edwards' estimates indicate large differences in the standard error of the regression (*SEE*) of the predictor of money growth and some deviations from Lucas' hypothesis. In particular, Peru has a relatively large *SEE* in the monetary prediction equation, but has one of the largest and most significant coefficients of unexpected money growth in the output equation.[3] It is only when I used past inflation rates (December to December) as regressors to predict money growth in Brazil and Chile that I obtained first, a positive (though still insignificant) relationship between unexpected money and output growth in these two countries, and second, similar *SEE*s for the monetary predictor and coefficients in the five countries which did not differ significantly from each other or from .1. (See my Table 3.) Both these results differ from Edwards' because he uses what is basically an autoregressive predictor of money, rather than one based on past inflation.

The use of past inflation rates as regressors to obtain predictions of money growth can be rationalized in many ways. (See for example, Thomas Sargent and Neil Wallace.) In addition, as a standard of comparison, I tried a simplified (constrained) version of this approach, using only the well-known proxy of lagged inflation as the predictor, and obtained relatively good results. However, as mentioned in the earlier article, it is important to use the December-to-December price changes in the two high-inflation countries.[4] Motivated by Edwards' questioning of these results I have rerun the equations, using *IFS* output data for Brazil to make future comparisons easier. (See Table 1.) As Edwards points out in his comment, the coefficient of *DM* is not significant, but, as opposed his results, the hypothesis of equal but opposite signs for the coefficients, cannot be rejected in either case.[5] Thus our differences in data—either the *GDP* or the price series —are of some importance here.

To summarize, I expected that my 10 percent rule of thumb for the response of output to unexpected inflation would suffer the fate of most empirical propositions in social science and be overturned when better predictors of inflation were found. However, I felt that Lucas' proposition of an inverse relation between this response and the predictability of inflation was on more solid ground. Edwards has made a contribution by (a) estimating the magnitude of the contribution of the deficit to monetary growth, and (b) showing that the terms of trade may influence output. However, I do not feel he has overturned my rule of thumb because his

---

[3] Edwards confuses predictability (measured, for example, by the *SEE*) with variance in money growth, in arguing that his reports support Lucas' hypothesis of the inverse relationship between unexpected inflation and the response of the economy to unexpected monetary growth.

[4] Moreover, one cannot directly use data from the *IFS* tapes for the early years in Brazil and Chile, since the data are reported to only one significant decimal place because of the high rate of inflation.

[5] The $F$-statistics are 2.9 (1,20) and .33 (1,16) respectively, versus critical values of 4.35 and 4.45. Note that the power of the test is low because the coefficient of *DM* is not significantly different from zero.

TABLE 1—OUTPUT GROWTH ($DY$) AS A FUNCTION OF MONEY GROWTH AND LAGGED INFLATION

| Equation | | Constant | $DM_t$ | $DP_{t-1}$ | $DM_t - DP_{t-1}$ | $R^2$ | D-W |
|---|---|---|---|---|---|---|---|
| (1) | Brazil | .088 | .094 | −.180 | — | .22 | 1.77 |
| | | (5.1) | (1.1) | (2.4) | | | |
| (2) | | .061 | — | — | .168 | .14 | 1.52 |
| | | (7.7) | | | (2.2) | | |
| (3) | Chile | .040 | .074 | −.104 | | .26 | 1.69 |
| | | (2.7) | (1.2) | (2.6) | | | |
| (4) | | .030 | — | — | .108 | .28 | 1.71 |
| | | (7.7) | | | (2.9) | | |

*Sources*: Brazil, *International Financial Statistics*, various issues and tape; Chile, Ricardo Ffrench-Davis.
*Notes:* The $t$-statistics are in parentheses. Equation (4) is reproduced from my earlier paper. $Y$ refers to $GDP$, $M$ to $M1$, $P$ to wholesale prices, and $D$ to the first difference of the *log* of the variable. $R^2$ is the multiple correlation coefficient adjusted for degrees of freedom.

monetary prediction equation using the fiscal deficit probably is not applicable in the countries analyzed, and, despite including this variable, our results are about the same in the medium-inflation countries while for the high-inflation countries, the residuals from my inflation-based prediction equations both outperform his autoregressive predictors in the output equation, and support the 10 percent rule.

## REFERENCES

**Barro, Robert,** "An Equilibrium Approach to Business Cycles," in his *Money, Expectations, and Business Cycles*, New York: Academic Press, 1981, 41–78.

**Edwards, Sebastian,** "The Short-Run Relation Between Growth and Inflation in Latin America: Comment," *American Economic Review*, June 1983, *73*, 477–82.

**Ffrench-Davis, Ricardo,** *Políticas Económicas en Chile, 1952–1970*, Santiago: Editorial Nueva Universidad, 1973.

**Hanson, James,** "The Short-Run Relation between Growth and Inflation in Latin America: A Quasi-Rational or Consistent Expectations Approach," *American Economic Review*, December 1980, *70*, 972–89.

**Sargent, Thomas and Wallace, Neil,** "Rational Expectations and the Dynamics of Hyperinflation," *International Economic Review*, June 1973, *14*, 328–50.

# Contestable Markets: An Uprising in the Theory of Industry Structure: Comment

*By* MARTIN L. WEITZMAN*

A situation traditionally identified as conducive to imperfect competition is when average costs decline and the cheapest scale of production is large relative to the size of the market. However, some of the more recent literature on industrial organization has emphasized a supposed distinction between *fixed* costs of production and *sunk* costs of production. A point frequently made is that while the possibility of sunk costs can create genuine barriers to entry, pure fixed costs per se do not inhibit competitive market performance. This approach finds its logical extreme in the norm or abstraction of "perfect contestability."

Perfect contestability is *essentially* a theory of the polar case of frictionless entry and exit. As William Baumol stated, "a contestable market is one into which entry is absolutely free and exit is absolutely costless" (p. 3). Certainly this is the most reasonable interpretation of a situation where potential entrants feel free to disregard an incumbent's price response.

Suppose every potential and actual producer has access to the same technology. In a perfectly contestable market, there are no set-up or shut-down losses. The main result is that even with declining average costs of production, the incumbent firm does not dare to post a price higher than average cost. Price above cost invites being undercut by a hit-and-run shadow entrant capable of producing at the same flow rate and at identical unit cost during just the briefest instant of time.

I argue that this line of reasoning is misleading. Perfect contestability gets around the problem of increasing returns only by, in effect, assuming it away. A hit-and-run technology makes the firm behave "as if" it is competitive in a market precisely because the convexity preconditions for competition are de facto being met in that market.

This comment shows that, strictly speaking, there is no such thing as a "pure" fixed cost. Unless there are sunk costs located somewhere in the relevant production technology, all costs are variable. As a matter of formal theory, you cannot have a range of decreasing average cost without sunk costs. Presumably there is also an approximation theorem which states that when sunk costs are "close" to being negligible then average production costs are "practically" nondecreasing. Perfect contestability holds approximately in a market only to the extent that production costs for the market are approximately nondecreasing.

My discussion will be restricted to the familiar case of a single, well-defined, homogeneous, fully divisible commodity. While more general approaches are possible, for the sake of simplicity the following definition is used.

The average cost function $AC(y,t)$ describes, say, the minimum cost per unit of output produced at a uniform flow rate $y$ throughout the time interval $(0,t]$. Especially in applications to industrial organization theory it is essential to remember that cost functions are not generally timeless, and that writing $AC$ as a function of $y$ alone can be dangerously vague.

Behind the free entry and exit of a hit-and-run technology in a perfectly contestable market is the abstraction of no sunk costs—investments are fully reversible because nothing is lost in setting up or shutting down production.

DEFINITION: *There are no sunk costs whenever there is some function* $f(y)$ *with* $f(0)=0$ *such that for all* $y$ *and* $t$, $AC(y,t)=f(y)$.

It might be conjectured that the absence of sunk costs does not preclude a declining section of the average cost curve. However, the following result shows that declining costs are not compatible with a timeless cost function.

---

*Massachusetts Institute of Technology

THEOREM: *If there are no sunk costs, the average cost function is nondecreasing over its entire range.*

PROOF:

Let $y < y'$. Pick any $t > 0$. Let $i = 1, 2, \ldots, n$. Consider the following production schedule. Throughout the time interval,

$$((i-1)t/n, (i-1)t/n + (y/y')t/n],$$

output is produced at flow rate $y'$. Throughout the time interval,

$$((i-1)t/n + (y/y')t/n, it/n],$$

output is produced at flow rate 0.

In the limit as $n$ is large, the production flow rate is uniformly $y$ throughout any subinterval of $(0, t]$. The cost per unit of output produced by this limiting schedule is $f(y')$. But then $f(y) \leqslant f(y')$.

My conclusion is that contestability cannot serve as a useful focal point for a general theory of market forms because it is meaningless to simultaneously postulate the presence of fixed costs and the absence of sunk costs, at least for any well-defined situation lending itself to the usual cost function approach. In the limit, as the concept of contestability is pushed hard, a hit-and-run technology cannot be other than convex.

## REFERENCE

**Baumol, William J.**, "Contestable Markets: An Uprising in the Theory of Industrial Structure," *American Economic Review*, March 1982, *72*, 1–15.

# Contestable Markets: An Uprising in the Theory of Industry Structure: Comment

*By* MARIUS SCHWARTZ AND ROBERT J. REYNOLDS*

In his AEA presidential address, William Baumol (1982) describes the theory of contestable markets as "an uprising" which "aspires to provide no less than a unifying theory as a foundation for the analysis of industrial organization" (pp. 14–15). A perfectly contestable market is defined as one into which entry is absolutely free and from which exit is absolutely costless (p. 3). In such a market, the price-taking results of the perfectly competitive model apply even if there is only a small number of incumbent firms. The crucial feature of a perfectly contestable market is its vulnerability to hit-and-run entry. A potential entrant could exploit even the most transient profit opportunity by instantaneously entering, collecting his profit, and exiting before incumbents could lower price: "...potential entrants find it appropriate to evaluate the profitability of entry in terms of incumbent firms' pre-entry prices" (p. 4). To prevent hit-and-run entry, incumbents must produce efficiently and earn zero profits. In a perfectly contestable market, therefore, the threat of entry suppresses any oligopolistic interaction that might occur among a small number of firms.

We will show that perfect contestability requires two implausible conditions:[1] (i) in response to high prices, an entrant can enter instantaneously at any scale, that is, there is no entry lag; and (ii) an entrant can undercut an incumbent's price and exit with no loss of fixed costs before the incumbent can adjust price, that is, the incumbent's price adjustment lag exceeds the exit lag.[2] If these conditions are relaxed even slightly, the results can differ dramatically from those obtained under perfect contestability. Specifically, if (ii) holds but (i) is violated, the incumbent may charge a high price pre-entry even though entry will eventually bring price down. If (ii) is violated, entry may not occur; whether it occurs depends upon the nature of the nondegenerate oligopolistic game expected post-entry.[3]

Consider a homogeneous good market and assume initially that there is a single incumbent and a single potential entrant. Suppose that (ii) above holds but (i) does not, that is, the incumbent's price-adjustment lag exceeds the exit lag but there is some entry lag. Because the entrant can enter at any scale, undercut the incumbent, and exit before the latter can adjust price, any supracompetitive pre-entry price (i.e., any price above the level yielding zero profit net of an allowance for interest on the fixed cost) will induce entry and subsequent loss of the incumbent's entire market. The incumbent will set his pre-entry price at one of the following two levels: either the competitive level or the monopoly level. The competitive price yields a zero profit stream forever and no entry. The monopoly price induces entry and some subsequent loss (since the incumbent is undercut and earns zero revenue but incurs an opportunity loss equal to the interest foregone on his fixed cost); however, pre-entry the incumbent earns monopoly profits. Depending on the relevant parameters, either of these pre-entry prices could maximize the incumbent's present value. Furthermore, if there is *no* exit lag—as in the perfectly contestable case—the choice is clear: since the incumbent can now exit and avoid any loss from being undercut post-entry, he will set pre-entry price at the monopoly level!

---

*Georgetown University, and U.S. Department of Justice and ICF Inc., respectively. The views expressed in this comment are not necessarily those of the U.S. Department of Justice or ICF. Helpful comments were received from Maxim Engers, Lucinda Lewis, Robert McGuckin and numerous colleagues at the Economic Policy Office of the Antitrust Division, U.S. Department of Justice.

[1] In interpreting perfect contestability we have relied in part on Baumol and Robert Willig (1981) and Baumol, John Panzar, and Willig (1982).

[2] We assume that all firms have the same cost function which exhibits some fixed, overhead cost. For analytical tractability, we represent the degree of difficulty in recovering the fixed cost (i.e., the notion of sunkedness) as the time it takes to leave the market with the fixed cost fully recovered. We call this time the "exit lag" and assume that it is the same for an incumbent as for an entrant. We also allow some lag between the time entry occurs and the time an incumbent can adjust price, and call this the "price-adjustment lag." Arguably, if the incumbent foresees entry he may be able to prepare in advance to either exit, or to adjust price when entry occurs, and may therefore be subject to no exit or price adjustment lags. In such a case, our ensuing criticisms of the nonrobustness of Baumol's results would be strengthened. By assuming that the incumbent faces both price adjustment and exit lags, and that the exit lag for the incumbent is the same as for the entrant, we are considering a case relatively favorable to Baumol's approach, It has been pointed out to us that Avinash Dixit (1982) also notes that conditions (i) and (ii) above are necessary for perfect contestability.

[3] There is a post-entry game in Baumol's model but it is degenerate, since the entrant can undercut the incumbent, gain the entire market and exit before the incumbent can adjust price.

In the above example, the threat of entry may not keep price low but actual entry will occur if the incumbent's pre-entry price is supracompetitive: the entrant can undercut the incumbent during the price-adjustment lag and exit before this lag elapses. Once we assume that the *exit* lag exceeds the price-adjustment lag, the entrant will be unable to escape price reactions by the incumbent and therefore may choose not to enter.

Consider the following example. Assume no price-adjustment lag and only arbitrarily small entry and exit lags. Conditions (i) and (ii) above are violated, but structurally the market is "almost perfectly contestable" as an arbitrarily small exit lag means that the fixed cost can be fully recovered almost immediately. However, the equilibrium is radically different from that in a perfect contestable market. Since price can now be adjusted instantaneously in response to entry, the entry decision now depends only on the nondegenerate oligopolistic game expected post-entry. Therefore, the incumbent will set pre-entry price at the monopoly level.[4] Moreover, the post-entry game may well imply negative profits to both firms (as in a Bertrand duopoly with identical firms having positive fixed cost and constant marginal cost), so the potential entrant may stay out despite the incumbent's high price.[5]

The assumptions of the above example can be relaxed in several respects without changing our qualitative results. First, there can be a positive price-adjustment lag; as long as this lag is shorter than the exit lag, the entrant will still have to consider the post-entry game in deciding whether to enter. Second, there can be several incumbents initially rather than one. The decision whether to enter will then depend upon a post-entry ologopolistic game involving several rather than two firms.[6] To summarize, Baumol's theory is not robust; once we deviate even slightly from the strict assumptions of perfect contestability, pricing and entry decisions depend upon the nature of firm interactions. There is no escaping the "reaction functions and the other paraphernalia of standard oligopoly models" (p. 11).

The nonrobustness of Baumol's results concerning industry behavior also calls into question whether the optimal industry structure (the size distribution of firms) said to prevail in perfectly contestable markets (pp. 5–6) is approximated in markets that are almost perfectly contestable. Baumol bases his structural results on the threat of hit-and-run entry which keeps prices and profits at competitive levels. We have seen that hit-

---

[4] In the first example, since the incumbent cannot change his price post-entry, he may set pre-entry price at the entry forestalling level. Here, the incumbent can lower price if entry occurs and therefore will set pre-entry price at the monopoly level.

[5] Characterizing the entrant's behavior in this situation is complex, because of the question of which firm would exit if the entrant were to come in. However, suppose that the entrant believed (with probability one) that the incumbent would exit. For certain parameter values, the losses that the entrant would incur while the incumbent remained in the market—because of the exit lag—would still exceed the present value of profit in the post-exit period, hence the entrant would stay out.

[6] If there are multiple *entrants*, the problem becomes more complicated since an entrant must know the number of such entrants, and whether they would enter sequentially or simultaneously. With a large number of entrants moving simultaneously, it is possible that none would enter because of the losses that would be incurred if all entered simultaneously (see, for example, Roger Sherman and Thomas Willett, 1967). The implications of different numbers of entrants and different interactions among them have not been thoroughly analyzed to our knowledge.

and-run entry may be impossible in markets that are almost perfectly contestable. Consequently, Baumol's conjecture (pp. 7–8) that equilibrium industry structure in almost perfectly contestable markets approximates that of perfectly contestable markets remains to be proved.

If the perfectly contestable market were a plausible case, it might not matter that Baumol's results do not extend beyond this case. However, as shown earlier, perfect contestability requires that an outsider can enter, produce, sell, and exit without loss—all before an incumbent can adjust price. Typical set-up costs and specificity of capital make this description grossly unrealistic. Indeed, it is precisely because incumbents can adjust price and output rapidly in response to entry that the early limit pricing models of entry deterrence (for example, Joe Bain), were criticized and replaced by subsequent theoretical work (for example, Earl Thompson and Roger Faith; A. Michael Spence; Steven Salop; Avinash Dixit, 1980; Reynolds and Salop; Paul Milgrom and John Roberts).

One final point should be noted. In some markets there may exist outside institutions that prevent price reactions by incumbents or negate their effect on entrants. For example, long-term contracts may enable entrants to contract for a fixed price before entering —as in Harold Demsetz's "competition-for-the-market" model.[7] Such contracts move us to an environment radically different from the usual oligopoly one. In this different environment, the threat of entry would ensure zero profit and approximately optimal performance even if irrecoverability of fixed costs precluded hit-and-run entry. The question of exit and therefore of recoverability of fixed costs would not arise, so the nature of equilibrium would be independent of whether the market was contestable in a structural sense.

---

[7] Such long-term contracts are likely to be infeasible because of the cost of monitoring and enforcing quality and performance, for example, Oliver Williamson, (1976). Long-term contracts are particularly unlikely where new entrants are involved since the quality and cost of their product is relatively unknown.

## REFERENCES

**Bain, Joe S.,** *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.

**Baumol, William J.,** "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, March 1982, *72*, 1–15.

______ **and Willig, Robert D.,** "Fixed Costs, Sunk Costs, Entry Barriers, and Sustainability of Monopoly," *Quarterly Journal of Economics*, August 1981, *96*, 405–31.

______, **Panzar, John C. and Willig, Robert D.,** *Contestable Markets and the Theory of Industry Structure*, San Diego: Harcourt Brace Jovanovich, 1982.

**Demsetz, Harold,** "Why Regulate Utilities?" *Journal of Law and Economics*, April 1968, *11*, 55–65.

**Dixit, Avinash,** "The Role of Investment in Entry Deterrence," *Economic Journal*, March 1980, *90*, 95–106.

______, "Recent Developments in Oligopoly Theory," *American Economic Review Proceedings*, May 1982, *72*, 12–17.

**Milgrom, Paul and Roberts, John.,** "Limit Pricing, Entry and Incomplete Information: An Equilibrium Analysis," *Econometrica*, March 1982, *50*, 443–59.

**Reynolds, Robert J. and Salop, Steven,** "Credible Limit Pricing and Entry," Working Paper, 1978.

**Salop, Steven,** "Strategic Entry Deterrence," *American Economic Review Proceedings*, May 1979, *69*, 335–38.

**Sherman, Roger and Willett, Thomas D.,** "Potential Entrants Discourage Entry," *Journal of Political Economy*, August 1967, *75*, 400–03.

**Spence, A. Michael,** "Entry, Capacity, Investment, and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, *8*, 534–44.

**Thompson, Earl A. and Faith, Roger L.,** "A Model of Non-Competitive Interdependence and Anti-Monopoly Laws," Working Paper No. 56, University of California-Los Angeles, 1974.

**Williamson, Oliver E.,** "Franchise Bidding for Natural Monopolies in General with Respect to CATV," *Bell Journal of Economics*, Spring 1976, *7*, 73–104.

# Contestable Markets: An Uprising in the Theory of Industry Structure: Reply

*By* WILLIAM J. BAUMOL, JOHN C. PANZAR, AND ROBERT D. WILLIG*

While the issues raised by Marius Schwartz and Robert Reynolds are quite different from those raised by Martin Weitzman, they have three significant common elements: an attempt to impose a dynamic mechanism upon a static equilibrium analysis, the provision of new and valuable insights, and extreme exaggeration of the generality of their results and their damage to contestability analysis. We will show here that, rather than destroying the theory, the insights of these able commentators enrich the analysis and open new areas for pertinent research. We will first discuss Weitzman's remarks and then turn to Schwartz and Reynolds.

## I. The Weitzman Parable

Weitzman denies that, in the absence of sunk costs, contestability theory is applicable to cases of natural monopoly, "natural oligopoly," and other imperfectly competitive arenas. His view is that perfect contestability is defined to require absolutely costless and instantaneous entry and exit, but that this in turn implies that returns to scale must be absolutely constant so that, for all practical purposes, perfectly contestable markets must all be perfectly competitive. Then, in his view, the theory is not really applicable to other market forms. To show why this conclusion is unwarranted, we first reformulate the Weitzman argument for the sake of clarity. Then we will show why it does not always work, and why it does not quite come to grips with the concept of contestability, properly defined. Finally, we will indicate the nature of the insight his comment does contribute.

To clarify the argument, let us translate Weitzman's few lines of algebra into a concrete example. Imagine a machine which operates at minimum average cost when it produces 12,000 units of output per hour. Suppose the machine can be moved instantaneously and costlessly from market to market. Then an entrepreneur who wishes to produce only 3,000 units of output per hour in each of four markets, A, B, C, and D, can shift the machine four times every hour, letting the machine run 15 minutes in each market. Thus, in market A, for example, between 1:00 and 1:15 P.M. it will produce its desired 3,000 units of output operating at its cost-minimizing rate of 12,000 units per hour. There obviously are no diseconomies to small-scale production here because of the perfect costlessness of entry and withdrawal.

Weitzman adds a final twist to this story in order to perfect his case. Suppose, for example, customers come into each market in a steady stream and do not want to wait for their purchases. Then by 1:15 our producer would have sold only 3,000/4 units of that hour's output, leaving (3/4)3,000 units in inventory. If inventory carrying costs are positive, this still imposes some higher costs upon the firm whose demand flow in that market falls short of the cost-minimizing rate of production of the machine. But Weitzman shows the way around this problem too. With exit and entry costless, the machine can just as well be shifted among markets not just four, but eight times per hour, each time operating only $7\frac{1}{2}$ minutes and producing 1,500 units of output in this interval. This time, the machine enters market A at 1:00 P.M., leaves it at 1:07$\frac{1}{2}$, reenters market A at 1:30, and leaves it again at 1:37$\frac{1}{2}$. By 1:07$\frac{1}{2}$, the producer will have already sold 3,000/8 units, so that unsold inventory will be only (3/4)(1,500) units, rather than (3/4)(3,000), as before. Nor does our entrepreneur have to stop there. He can increase the frequency of his intermarket moves to 16, or 32, or 64 per hour..., each time cutting his average inventory in half, and, in the limit, eliminating inventory completely. In this way, all financial penalties for small-scale production can be eliminated and so, by virtue of costless entry and exit, a universal regime of constant returns to scale is guaranteed. Finally, note that the costless physical mobility of a machine from market to market in this parable can be replaced by the existence of a rental market in which the entrepreneur can obtain the services of the machine for an arbitrarily short period of time without paying a higher rate per unit of time.

---

*Princeton and New York Universities, Bell Laboratories, and Princeton University, respectively. The views expressed are our own and do not necessarily reflect those of Bell Laboratories or the Bell System.

This, then, is the Weitzman story. Why does it not hold generally?

## II. Perishability, Batch Production, and Time Requirements

Many outputs and production processes may fit Weitzman's parable as a limiting case, but many do not. First, some items, notably services, cannot be stored at all. They perish the instant they are produced, and one cannot get around this problem by taking the time they are held in inventory to be very brief, approaching zero in the limit. Second, fixed investment does not always take the form of capacity capable of turning out a continuous steady flow of output. Rather, some types of processes require an irreducible amount of production *time* to be effective and in that time they yield some fixed minimum batch of output. More generally, the minimal average cost per unit produced may be achievable only if the production process is run at a particular intensity over a particular span of time. In these cases, Weitzman's technique of "substituting" the perfect divisibility of time to compensate for other indivisibilities is inapplicable.

The relevance of these two phenomena, which expressed as abstractions may appear to be quibbles, is brought out by another example. Along any airline route, say Philadelphia to Ithaca, cost per passenger is minimized by a jumbo jet if it is used to capacity, some 400 seats. Suppose no more than 100 passengers travel this route every day. Even if airplanes could be switched from one route to another without cost or delay, there is no way the Weitzman parable can be stretched to fit. Transportation services cannot be stored. There is no way any of the capacity of the departed plane can be preserved for those passengers who wish to begin their trips after it leaves.

More important, the airplane inherently provides its services over time in one 400-seat batch. Since production of the service requires some nonvanishing interval of time, say one hour, there is no way in which an abbreviated period of operation on the Philadelphia-Ithaca route can lead to the achievement of optimal scale. One simply cannot do it by operating a 400-seat plane carrying 100 people for 15 minutes. Of course, it can be flown only once every four days, but that technique supplies a different and inferior product which a smaller and less economical plane may be able to outcompete with a daily schedule. Thus, when production via the most efficient techniques occurs in minimum batches (400 seats per plane, 12,000 shoes per production run) and the size of the batches cannot be cut by abbreviating the time employed in the process, the Weitzman argument clearly breaks down.

Since there are many goods for which batch production is most efficient, there is simply no validity to the view that the Weitzman argument applies to all, or even most, manufacturing. Perfect contestability as a substantive generalization of perfect competition can apply to manufacturing just as well as to the provision of services.

We conclude that, even as a limiting case, the Weitzman parable is able to generate constant returns to scale only for products which are storable, at least briefly, and which do not require an indivisible "lump" of time for their most efficient batch production. Neither class of goods—those which fit Weitzman's parable and those which do not—is economically unimportant.

## III. The Meaning of "Sunk Cost"[1]

Weitzman's analysis suffers from another more fundamental problem—a mischaracterization of contestability resulting from a confusion between *economic* and technological notions of sunk costs. To produce its results, even the limiting case of perfect contestability does *not* require entry and exit to be instantaneous. Rather, it is sufficient that the process be rapid enough so that the entrant does not find his investment vulnerable to a retaliatory response by the incumbent. *The length of this time period is not exclusively a technological datum, but is also the result of business practice and opportunities in the market in question.* This period can be as long as the longest period for which it is credible for buyers to commit their patronage to the entrant. In terms of an airline scenario, "hit-and-run" entry may be possible because the entrant can sell tickets *conditionally, and before* he flies his plane into the market, not because he can fly it in faster than the incumbent is able to cut his price. For example, before deregulation it was a common practice of charter airlines to make contracts (ticket sales) conditional on filling the plane.

---

[1] We would like to thank Sanford Grossman for helping to clarify our thoughts in this area.

With contracting feasible, the fact that successful entry may require commitment of assets to a particular market for a nontrivial interval of time need not diminish the viability of hit-and-run entry, nor imply the presence of costs which are sunk in any economically significant sense. Even if such contracting is not feasible, it is still possible for regulation, costs of communicating price revisions, or other impediments to delay an incumbent's effective price response for a period of length $T > 0$. In either case, the concept of hit-and-run entry can survive the technological imposition of a minimum production-time requirement $t^* \leqslant T$, while the Weitzman parable cannot. Thus, it is far from "meaningless" to posit the presence of fixed costs in the absence of sunk costs.

## IV. Toward Theory of "Almost Contestable" Markets

Schwartz and Reynolds raise an issue that is very different and, in our view, considerably more significant. They point out, correctly, that sufficient conditions for *perfect* contestability are virtually certain to be violated in reality, at least to some degree. They ask whether small deviations from the conditions they cite are likely to affect substantially the conclusions of the analysis. They argue that there is, indeed, a sharp discontinuity, so that minor violations of the conditions they cite are likely to cause large losses in the social benefits promised by perfect contestability.

We feel that the issue they raise is an important subject for analysis. We will show, however, that their model of deviation from perfect contestability is only one of the pertinent constructs that is possible; that other plausible and legitimate constructs imply that behavior and social benefits approach those of perfect contestability continuously as the contestability of markets increases; and that both intuition and preliminary empirical evidence suggest the hypothesis that, in the world of reality, a reduction of barriers to entry and exit, particularly those taking the form of sunk costs, yield steady improvements in industry performance and behavior, as robustness of the policy implications of contestability analysis requires.

Let us first consider the construction of a systematic analysis of the effect upon a market of the introduction of a small amount of sunk costs. This analysis should not base itself on the rather special premises adopted by Schwartz and Reynolds, but rather on the economically relevant progression from the limiting case in which entry does not subject any capital to the risk of retaliatory responses by the incumbent (complete absence of sunk costs) to one in which a small amount of an entrant's investment must be sunk irrevocably in the market. We show that, under plausible assumptions, the associated change in the equilibrium price must also be small. All we require for this result is that, before making his investment in the market, the potential entrant be able to enter into contracts to supply potential customers for a nontrivial interval of time, $T$. Schwartz and Reynolds discuss this possibility briefly in their last paragraph, but by focusing only on *long-term* contracts, they fail to realize that this *is* the relevant point of departure.

The basic scenario is quite simple. At time zero, the potential entrant may offer to serve customers for a period of time $T$ at a price $p_e \leqslant p_i$, the price of the incumbents. Buyers would, quite naturally, inquire whether any incumbent intended to match $p_e$. If one did, then entry would not occur, *but the price $p_e$ would prevail in the market for T units of time*, whereupon an entrant can again appear with an offer waiting to be matched. Thus, in the absence of exogenous change, $p_e$ can be expected to prevail in the market at all times. How low a $p_e$ will a potential entrant find it desirable to offer? This depends upon the magnitude of the *economically* sunk costs. To permit an analysis of these, suppose that a unit of capital purchased at a price of $\beta$ per unit could be sold or utilized elsewhere after $T$ for a unit salvage value of $\alpha \leqslant \beta$. Thus, it is possible to parameterize continuously the degree of sunkenness of capital from zero ($\alpha = \beta$) to absolute sunkenness ($\alpha = 0$). If $\alpha = \beta$, the potential entrant can afford to offer a price just equal to the economically efficient level of average cost, because its flow of sales is guaranteed until $T$, and afterwards it has no "unamortized capital" to be concerned about. However, if $\alpha < \beta$, the entrant can offer a price only as low as *its* average cost over $T$, and this will be somewhat higher than the true economic average cost because some of the economic value of its plant may be lost as a result of exit or of oligopolistic rivalry which takes place after the initial contracts expire at time $T$. Then, incumbents can charge a price above true economic average cost without inducing entry, because of the sunk economic costs required of entrants. The sustainable elevation of price, here, is larger the smaller is $\alpha$. And $\alpha$ will be smaller the larger the proportion of cost that is technologically sunk, and the shorter is the time period, $T$.[2] Thus, it is clear that the deleterious effects of sunk costs behave "continuously" in this framework, in contrast to their behavior in the analysis of Reynolds and Schwartz.

The essential difference between our scenario and that of Reynolds and Schwartz is that buyers' behavior is assigned a clear role in our version of the entry process. Do buyers have an incentive to enter into contracts in the way we describe? In a real sense they do, for, *ceteris paribus*, the longer is $T$, the lower is the price that will be available from entrants and the tighter is the constraint that potential entry places on the incumbents' pricing policy. Of course, there may be influences operating in the opposite direction, such as the desire for flexibility in the face of uncertainty, etc. The construction of explicit models of buyers' behavior may well be the most promising direction for further research in this area.

The moral of all this is that models of behavior under imperfect contestability other than that of Schwartz and Reynolds are possible and can be extremely plausible.[3] In the model discussed above, and no doubt in others as well, where there are almost no sunk costs, markets are almost perfectly contestable.

It should be emphasized that models which support the robustness of contestability analysis follow a relatively long tradition going back at least to the work of J. S. Bain. This tradition holds that increased ease of entry and exit improves the welfare performance of firms and industries. On this subject, the theory of contestability has only sought to contribute insights on the underpinnings of that judgment. Certainly it is a view widely accepted on the basis of casual observation, though systematic empirical testing still remains to be carried out. A first step in this direction is the recent empirical work by Ioannis Kessides (1982), which shows that among 4-digit U.S. manufacturing industries, entry increases with the profit levels of incumbents, holding fixed the expected losses from the possibility of exit. These expected losses are greater, Kessides finds, the greater are the sunk costs (as measured by several proxies) that must be committed by an entrant. Sunk costs deter entry and also diminish the rate at which entry responds to incumbents' profits. In Kessides' estimated model, these effects are statistically significant and they behave in a continuous and monotonic manner. They lend some support to the hypothesis that market performance depends continuously on the degree of imperfection in their contestability.

---

[2] A model with these features and results is described and analyzed in our 1982 book, ch. 10; and in Section II of our forthcoming paper.

[3] Another example is provided by recent theoretical work by Mark Bagnoli, a graduate student at Princeton University. This focuses on the effects of the advertising costs that must be incurred by firms to inform potential customers about an initial posted price and about any subsequent price changes. Where such outlays are sunk, but no other costs are sunk, incumbents are strictly motivated to adopt sustainable (zero profit) prices. For if they do not, entry will occur and they will only be able to respond via a price reduction whose advertising cost makes the response a losing proposition. The best response of the incumbent whose initial prices were unsustainably high is to exit when entry occurs, thereby losing earlier advertising outlays.

## V. Perfect Contestability as a Static Equilibrium Analysis

There is another critical misunderstanding that characterizes both comments on contestability theory, despite the great difference in their themes. For both seek to criticize what is essentially a static equilibrium theory by *imputing to it* a dynamic mechanism *of their own creation*, and attacking the reasonableness of the assumptions *they* have made about the dynamic process governing the behavior of agents in contestable markets outside of equilibrium. This is not to say that dynamic considerations are unimportant, rather that they should not be inferred casually from an informal statement of a model's equilibrium properties.

The materials to which we now turn may at first glance seem excessively formalistic, but they bring out the fact that most of the theorems and results which have been derived from contestability theory rest exclusively on static analysis and are in no way dependent on the dynamic construct imposed upon it either to give it credence or to attack its applicability. Let us, then, provide a brief and relatively formal definition of a perfectly contestable market, which is only hinted at in the presidential address:

DEFINITION 1: *A feasible industry configuration is composed of m firms producing output vectors* $y^1,\ldots,y^m \in R^n_+$ *at prices* $p \in R^n_{++}$, *such that markets clear,* $\sum_{i=1}^m y^i = Q(p)$, *and that each firm at least breaks even,* $p \cdot y^i - C(y^i) \geqslant 0$, $i = 1,\ldots,m$, *where* $Q(p)$ *is the market-demand relationship and* $C(y)$ *is the total cost function of each firm.*

DEFINITION 2: *A feasible industry configuration with price p and firms' output vectors* $y^1,\ldots,y^m$ *is sustainable iff* $p^e \cdot y^e \leqslant C(y^e)$ *for all* $p^e \leqslant p$ *and* $y^e \leqslant Q(p^e)$.

DEFINITION 3: *In a perfectly contestable market, an equilibrium industry configuration must be sustainable.*

Definition 2 provides a precise analytical statement of some of the requirements for a *static equilibrium* under free entry whenever "potential entrants find it appropriate to evaluate the profitability of entry in terms of the incumbent firms' pre-entry prices" (Baumol, p. 4). The theory of perfectly contestable markets, then, is neither more nor less than the derivation of the logical implications of this concept of multimarket (partial) equilibrium. While the "theory" is still young and its empirical implications largely untested, it has some fundamental properties which we think recommend it as a benchmark, a limiting case for the study of industrial organization. In particular, it is a substantive *generalization* of the traditional theory of perfect competition. This can be substantiated via a formal definition of long-run competitive equilibrium:

DEFINITION 4: *A feasible industry configuration (Definition* 1) $p; y^1,\ldots,y^m$ *is a long-run competitive equilibrium iff* $p \cdot y \leqslant C(y) \forall y \in R^n_+$.

So defined, a long-run competitive equilibrium has all its traditional properties. Each firm in the configuration takes prices as parametric, chooses output to maximize profit, earns zero profit, and equates marginal costs of goods produced to their prices. We can now state

PROPOSITION 1: (i) *A long-run competitive equilibrium is a sustainable industry configuration, so that perfectly competitive markets are always perfectly contestable.* (ii) *However, the converse is not true.*

Part (i) follows easily from the definitions. The example of average cost pricing in the presence of increasing returns to scale (the

Demsetz solution) establishes (ii), for the lowest price at which the market demand curve intersects a falling average cost curve, together with single-firm supply of the items demanded by the market, is a sustainable industry configuration, but is certainly not a competitive equilibrium.

Notice that, contrary to the perceptions of Schwartz and Reynolds, the static equilibrium theory of perfectly contestable markets contains *no assumptions whatever* about the possibility of instantaneous entry or the ability of new firms to exit more quickly than an incumbent can change its prices. Such dynamic properties are best thought of as attempts to examine the plausibility and range of applicability of the static equilibrium concepts. Conditions (i) and (ii) of Schwartz and Reynolds are *not* required by perfect contestability, nor, as we have seen, are they a logical consequence of the absence of *economically* sunk costs, the attribute of a market which, we believe, makes the perfect contestability paradigm applicable. In short, while the conditions which Schwartz and Reynolds cite may be *sufficient* to ensure perfect contestability, they are certainly not *necessary*.

None of this is meant to deny that the issue of "robustness' which they raise is of significance to any theory, and we have sought to examine it in terms of the appropriate question: can markets in which there are "almost" no sunk costs behave "almost" perfectly contestably? We have shown that they can, and we believe empirical evidence will confirm that they generally do. In other words, our hypothesis, which remains to be tested empirically, is that the predictions of static contestability theory remain reasonably accurate when sunk costs are present but insignificant.

### VI. Concluding Comments

It is important to end the discussion constructively. Despite the unavoidable emphasis in the preceding pages on our disagreements with our critics, we believe that they have provided important insights and made significant contributions to the theory. Schwartz and Reynolds have rightly called attention to the need for investigation of the consequences of small violations of the premises of perfect contestability, and have provided illuminating models of the effects of lags in entry, exit, and price adjustment upon the behavior of firms. Weitzman has revealed that there is an (apparently important) class of cases in which ease of entry and exit may be associated with little or no economies of scale. In such cases, the absence of entry barriers may derive part of its power to restrain monopoly from its ability to make the cost relationship approximate constant returns to scale more closely. Those are valuable insights which we believe will provide significant subjects for further research and enrich contestability theory and the theory of industry structure more generally.

## REFERENCES

**Baumol, William J.**, "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, March 1982, *72*, 1–15.

______, **Panzar, John C. and Willig, Robert D.**, *Contestable Markets and the Theory of Industry Structure*, San Diego: Harcourt Brace Jovanovich, 1982.

______, ______, **and** ______, "On the Theory of Perfectly Contestable Markets," in *New Developments in the Theory of Industry Structure*, International Economic Association: forthcoming.

**Kessides, Ioannis**, "Toward a Testable Model of Entry: A Study of the U.S. Manufacturing Industries, unpublished, Princeton University, June 1982.

**Schwartz, Marius and Reynolds, Robert J.**, "Contestable Markets: An Uprising in the Theory of Industry Structure: Comment," *American Economic Review*, June 1983, *73*, 488–900.

**Weitzman, Martin L.**, "Contestable Markets: An Uprising in the Theory of Industry Structure: Comment," *American Economic Review*, June 1983, *73*, 486–87.

# The Choice of Discount Rates for Public Projects: Comment

*By* JANDHYALA L. SHARMA*

In a recent article in this *Review*, Robert Mendelsohn attempted to clarify David Bradford's misunderstanding regarding optimal fraction of capital ($S_k$) that is consumed and fraction of income that is reinvested ($S_y$). One of the criteria suggested for selecting a public projects is the present value of the future consumption income from the investment, as stated by Mendelsohn: "The present value of dollar of investment is worth even more if future returns from that investment are reinvested" (p. 240). The social present value of dollar of private investment, $V_t$, varies considerably with the parameters, the market rate of interest, $r$, the social rate of time preference, $i$, and the rate of reinvestment, $S_y$. However, the impact of $S_y$ on the present value, $V_t$, and its role in the selection process is not clear. The variable $S_y$ is rather ambiguous and the optimum value, in general, for $S_y$ is either zero or one. The formula mentioned at the bottom of Mendelsohn's Table 1 (p. 240) should read:

$$(1) \qquad V_t = (1 - S_y)r/i - S_y r.$$

The present value criteria for deciding public projects is riddled with many problems. The impact of marginal propensity to save, $S_y$, is rather a passive variable. This makes the present value concept inappropriate for decision making. The behavior of $V_t$ is examined for different values of $S_y$, given $i$ and $r$; and $0 \leqslant S_y < 1$ and $V_t$ is given by equation (1).

*Case* I: $i > r$. The various values $V_t$ can take $V_t = r/i$, if $S_y = 0$, and $V_t = 0$, if $S_y = 1$. Since the values of $V_t$ are declining at an increasing rate ($V_t' < 0$, $V_t'' < 0$), the optimum value of $S_y$ is either 0 or 1, and so no unique value.

*Case* II: $i = r$. The value $V_t$ is equal to $i/r$, for all values of $S_y$; again yielding no unique value. Hence, $V_t$ is independent of $S_y$ and all values of $S_y$ are equally optimal.

*Case* III: First, equation (1) is derived under the assumption $i > r$ (strictly speaking, $i > S_y r$) because $i > r$ is a necessary condition of present value series convergence as $t \to \infty$; that is, the market rate of return ($r$) should be less than the social rate of time preference ($i$) in order to derive $V_r$, the present value of dollar investment earning $r$ percent. If $r > i$, then the series is an exploding series and $V_t \to \infty$, as $t \to \infty$. Hence, $V_t$ has no finite value. However, the behavior of $V_t$ can be studied under conditions $i < r$, and $V(t)$ is given by equation (1): $V_t = r/i$, if $S_y = 0$, and $V_t \to \infty$ as $S_y \to i/r$. So, if $i < r$, there is some value of $S_y$, to wit, $i/r$ which drives $V_t$ to infinity and finite for all values of $S_y < i/r$. As before, $V_t = 0$ if $S_y = 0$, since the function ($V_t$) is upward sloping, given $S_y = i/r$, and increasing at an increasing rate ($V_t' > 0, V_t'' > 0$), no unique values are assured. Optimal value of $S_y$ is either 0 or $i/r$.

Thus, the reinvestment rate ($S_y$) has no optimum value. The optimum values are either $S_y = 0$ or 1, so $S_y$ should not play any critical role in deciding public projects. Thus, the solution to the problem of choosing appropriate values of $S_y$, $i$, and $r$ is as elusive as ever and savings rates play only a passive role. Since $S_y$ is under decision maker's control, the present values can be varied, depending on the value assigned to $S_y$, the marginal propensity to save. The crucial issue in deciding the public projects is the determination of the social rate of time preferences and not the savings rates. The present value of public projects can be controlled by choosing arbitrary values for marginal propensity to save.

## REFERENCES

**Bradford, David,** "Constraints on Government Investment Opportunities and Choice of Discount Rates," *American Economic Re-*

---

*Cleveland State University

*view*, December 1975, *65*, 87–99.

**Mendelsohn, Robert,** "The Choice of Discount Rates for Public Projects," *American Economic Review*, March 1981, *71*, 239–41.

**Marglin, Steven,** "The Social Rate of Discount and the Optimal Rate of Investment," *Quarterly Journal of Economics*, February 1963, *77*, 95–111.

# The Choice of Discount Rates for Public Projects: Reply

By ROBERT MENDELSOHN*

Jandhyala Sharma makes two interesting points in his comment on the choice of discount rates for public projects. First, if the market interest rate $r$ exceeds the social discount rate $i$, it is possible that my (1981) formula for calculating the consumption value of an investment dollar is inappropriate, and second, if the marginal propensity to save from income $S_y$ is a decision variable, its optimal value is 0 or 1. Sharma concludes that $S_y$ or reinvestment should not play a critical rôle in deciding among public projects.

Sharma correctly argues that if the present value of the consumption stream generated by a dollar of private investment does not converge, my formula is inappropriate. Contrary to Sharma's assertions, the necessary conditions for convergence are that $i > S_y r$ (not $i > r$). The correct formula for the present value of an investment dollar $V_t$ is[1]

$$V_t = (1 - S_y)r/(i - S_y r) \quad i > S_y r$$
$$V_t = \infty \quad i \leqslant S_y r.$$

In the Bradford-Mendelsohn model, all reinvestment decisions are made by private individuals, so that $S_y$ is exogenous to the public project. Although he fails to motivate this assumption, Sharma assumes that $S_y$ is under the control of the decision maker so that it is endogenous to the project. Let us imagine when $S_y$ could be endogenous. Suppose the marginal propensity to save from income by specific groups of individuals were known and were different. To the extent that the benefits or costs of the project could be redistributed among the groups, one could alter $S_y$ and thus the resulting reinvestment rates.

Suppose, then, that $S_y$ is endogenous. Because of taxes on capital income, $r > i$. Taking the derivative of $V_t$ with respect to $S_y$ yields

$$\partial V_t / \partial S_y = r(r - i)/(i - S_y r)^2$$
$$\text{when} \quad i > S_y r.$$

If $i > S_y r$, and $r > i$, then the partial derivative of $V_t$ with respect to $S_y$ is positive. When the value of $V_t$ is infinite, it is still true that $V_t$ increases with $S_y$ since the consumption stream tends to diverge more rapidly the greater is $S_y$. The optimal value of $S_y$ for benefits is therefore 1, and for costs it is 0. For otherwise identical projects, one would prefer the project with the highest reinvestment rates of benefits and the lowest reinvestment rates of costs (foregone income). Sharma's conclusion that $S_y$ has no role in choosing among public projects is therefore incorrect. Even when $S_y$ is endogenous, it is quite clear that $S_y$ has a critical role to play in the choice of public projects.

In general, little is known about the reinvestment rates of various groups, so that $S_y$ is often best treated as an exogenous parameter. One possible exception to this rule hinges upon the commonly presumed statement that $S_y$ is higher for upper-income individuals. If this is true, the more the direct benefits of a public project can be concentrated on higher-income individuals, and the costs upon the poor, the greater will be the net benefits of the project. For example, if the market interest rate is 10 percent, the social time preference is 5 percent, and the marginal propensities to save from income are 0 and .40, respectively, for the poor vs. rich, the social value of a dollar of cash benefits to the poor is \$1 while to the rich the same dollar is worth \$3.

Although society may not choose to maximize net benefits by aiding the rich at the expense of the poor, the analysis suggests that the true efficiency costs of achieving vertical equity through public projects may be far greater than originally perceived.

---

*Department of economics, University of Washington.

[1] Note that the formula for $V_t$ in Table 2 of my 1981 paper (p. 240) was originally misprinted.

## REFERENCES

**Bradford, David F.,** "Constraints on Government Investment Opportunities and the Choice of Discount Rate," *American Economic Review*, December 1975, *65*, 887–99.

**Mendelsohn, Robert,** "The Choice of Discount Rates for Public Projects," *American Economic Review*, March 1981, *71*, 239–41.

**Sharma, Jandhyala L.,** "The Choice of Discount Rates for Public Projects: Comment," June 1983, *73*, 497–98.

# Expected Inflation and Interest Rates: Comment

*By* VITO TANZI*

In attempting to explain why, up to recently, the level of interest rates had been much lower than one would expect from the modified version of Fisher's theory, that takes into account the effect of income taxes on interest income, Michael Melvin, in his 1982 article in this *Review*, criticizes my 1980 hypothesis of a fiscal illusion and advances as a plausible explanation the Mundell Effect. He states:

> ...[A]s expected inflation increases, an excess demand for goods is created along with an excess supply of money (as $i$ rises in response to the increase in $\pi$). As prices rise due to the goods demand, real balances fall causing a shift in the savings function, which then leads to a drop in the after-tax real interest rate in order to restore equilibrium in the goods market. [p. 843]

Therefore, he writes, "We should not conclude that since the coefficient on $\pi$ is less than unity, people suffer from fiscal illusion..." (p. 843). The same point is made by Milton Ezrati in his 1982 comment on my paper (p. 857). The issue I wish to raise is not whether the Mundell effect exists, but, rather, whether it could account by itself for the difference between the estimated and the expected coefficient of the inflationary expectation variable.

Let $i$ = the (annual) nominal rate of interest, $r$ = the (annual) real rate, $r^*$ = the after-tax (annual) real state, $\pi$ = the expected (annual) rate of inflation, and $T$ = the tax rate on interest incomes. From the Fisherian relationship we get

$$i_t = r_t + \beta\pi_t, \tag{1}$$

where $\beta = di/d\pi$.

By taking realistic ranges of values for $r$, $\pi$, and $\beta$, equation (1) can be solved. Once this is done, the after-tax real rate $r^*$ can be obtained as follows:

$$r_t^* = i_t(1-T_t) - \pi_t. \tag{2}$$

The real rate $r$ will be assumed to equal 3. This is the estimate found in my earlier paper and is equal to, or even higher than, the estimates found by most other researchers. The expected inflation rate $\pi$ will be given values ranging from 2 to 20 percent per year to cover the full range experienced by the U.S. economy in recent decades. The $\beta$ is given values ranging from 0.70 to 1.00 to cover the estimates found in most empirical studies of the pre-1979 period. Melvin's own estimate of $\beta$, at 0.55, is even lower. The tax rate $T$ is given values ranging from 0.20 to 0.50 reflecting likely levels for this variable. The results are given in Table 1.

The Mundell effect maintains that inflation reduces the real rate. It surely does not argue that it makes it negative. However, Table 1 shows how quickly the net-of-tax real rate is reduced to zero, or even to negative values, when the nominal rate of interest does not adjust for the effect of taxes. Take, for example, the situation where $\beta = 0.70$ which has been, perhaps, the most typical coefficient found in empirical studies. At a tax rate of 30 percent, the after-tax real interest rate is already negative at a rate of inflation of 6 percent. By the time the rate of inflation reaches 12 percent, the after-tax real rate is a negative 4 percent. And, of course, increasing the tax rate or the rate of inflation, even lower values for $r^*$ are found. If the value of $\beta$ should be as small as in Melvin's estimate (0.55), the net-of-tax rate of return would be even smaller.

Clearly, it seems highly improbable that the Mundell effect, even if it exists, could explain more than a small part of the difference between the estimated and the expected coefficient of the inflationary expectation variable in the Fisher equation. If that effect provided the full explanation, so that no fiscal illusion were present, it would imply an implausibly high elasticity of the real

---

*Director, Fiscal Affairs Department, International Monetary Fund.

TABLE 1—SIMULATIONS OF AFTER-TAX EXPECTED INTEREST RATES

| $\pi$ | $T2 = 0.2$ | $T3 = 0.3$ | $T4 = 0.4$ | $T5 = 0.5$ |
|---|---|---|---|---|
| $\beta = 0.70$ | | | | |
| 2 | 1.52 | 1.08 | 0.64 | 0.20 |
| 4 | 0.64 | 0.06 | −0.52 | −1.10 |
| 6 | −0.24 | −0.96 | −1.68 | −2.40 |
| 8 | −1.12 | −1.98 | −2.84 | −3.70 |
| 10 | −2.00 | −3.00 | −4.00 | −5.00 |
| 12 | −2.88 | −4.02 | −5.16 | −6.30 |
| 14 | −3.76 | −5.04 | −6.32 | −7.60 |
| 16 | −4.64 | −6.06 | −7.48 | −8.90 |
| 18 | −5.52 | −7.08 | −8.64 | −10.20 |
| 20 | −6.40 | −8.10 | −9.80 | −11.50 |
| $\beta = 0.80$ | | | | |
| 2 | 1.68 | 1.22 | 0.76 | 0.30 |
| 4 | 0.96 | 0.34 | −0.28 | −0.90 |
| 6 | 0.24 | −0.54 | −1.32 | −2.10 |
| 8 | −0.48 | −1.42 | −2.36 | −3.30 |
| 10 | −1.20 | −2.30 | −3.40 | −4.50 |
| 12 | −1.92 | −3.18 | −4.44 | −5.70 |
| 14 | −2.64 | −4.06 | −5.48 | −6.90 |
| 16 | −3.36 | −4.94 | −6.52 | −8.10 |
| 18 | −4.08 | −5.82 | −7.56 | −9.30 |
| 20 | −4.80 | −6.70 | −8.60 | −10.50 |
| $\beta = 0.90$ | | | | |
| 2 | 1.84 | 1.36 | 0.88 | 0.40 |
| 4 | 1.28 | 0.62 | −0.04 | −0.70 |
| 6 | 0.72 | −0.12 | −0.96 | −1.80 |
| 8 | 0.16 | −0.86 | −1.88 | −2.90 |
| 10 | −0.40 | −1.60 | −2.80 | −4.00 |
| 12 | −0.96 | −2.34 | −3.72 | −5.10 |
| 14 | −1.52 | −3.08 | −4.64 | −6.20 |
| 16 | −2.08 | −3.82 | −5.56 | −7.30 |
| 18 | −2.64 | −4.56 | −6.48 | −8.40 |
| 20 | −3.20 | −5.30 | −7.40 | −9.50 |
| $\beta = 1.00$ | | | | |
| 2 | 2.00 | 1.50 | 1.00 | 0.50 |
| 4 | 1.60 | 0.90 | 0.20 | −0.50 |
| 6 | 1.20 | 0.30 | −0.60 | −1.50 |
| 8 | 0.80 | −0.30 | −1.40 | −2.50 |
| 10 | 0.40 | −0.90 | −2.20 | −3.50 |
| 12 | 0 | −1.50 | −3.00 | −4.50 |
| 14 | −0.40 | −2.10 | −3.80 | −5.50 |
| 16 | −0.80 | −2.70 | −4.60 | −6.50 |
| 18 | −1.20 | −3.30 | −5.40 | −7.50 |
| 20 | −1.60 | −3.90 | −6.20 | −8.50 |

rate with respect to changes in saving. Perhaps a Mundell effect, *together with* a fiscal illusion, provides the best explanation of why interest rates did not adjust for the effect of taxes in the pre-1979 period.

## REFERENCES

**Ezrati, Milton J.**, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates: Comment," *American Economic Review*, September 1982, *72*, 854–57.

**Melvin, Michael**, "Expected Inflation, Taxation, and Interest Rates: The Delusion of Fiscal Illusion," *American Economic Review*, September 1982, *72*, 841–45.

**Mundell, Robert A.**, "Inflation and Real Interest," *Journal of Political Economy*, June 1963, *71*, 280–83.

**Tanzi, Vito**, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates," *American Economic Review*, March 1980, *70*, 12–21.

# Expected Inflation and Interest Rates: Reply

*By* MICHAEL MELVIN*

Vito Tanzi's comment on my earlier study (1982) is based upon his belief that the Mundell effect alone could not likely account for the *ex ante* negative real rates observed in his simulations. His suggestion is that "Perhaps a Mundell effect, *together with* a fiscal illusion, provides the best explanation" (p. 502). Nowhere did I argue that the Mundell effect was all-encompassing. The nominal interest rate is determined by many factors besides expected inflation ($\pi$). My reduced-form nominal interest rate equation includes inflation uncertainty, the current and lagged money stock, and the lagged price level as well as expected inflation as arguments. Thus, at any point in time, the nominal rate will shift with a movement in any of these variables. One way of looking at this is that holding $\pi$ constant, $i$ will vary with the other predetermined variables so that the real rate changes. Since Tanzi and previous researchers have used partial equilibrium analysis to highlight the relation between $\pi$ and $i$, we tend to forget that the level of the real rate depends upon all the predetermined variables of the system. To concentrate on the partial derivative of $i$ with respect to $\pi$, we must take as given some initial $i$, holding constant the other factors influencing $i$. Tanzi's simulations set the real rate of interest ($r$) initially equal to .03. Then by assuming different values for $di/d\pi$ and $\pi$, he generates implied values for the after-tax real interest rate ($r^*$). Since the after-tax real rate turns negative at quite low rates of $\pi$, Tanzi concludes that fiscal illusion is present as "interest rates did not adjust for the effect of taxes" (p. 502).

The simple fact that the after-tax real rate turns negative need imply nothing regarding whether or not investors are cognizant of taxes. There is nothing inherently unusual about negative real rates of interest in the short run during inflationary periods. Clearly, if money can be costlessly stored, the nominal rate of interest cannot be negative but this in no way rules out negative real rates.

In a world of incomplete markets, there may be inflationary periods when a negative real rate of interest is the best that optimizing agents can earn. If there was a suitable asset available that guaranteed a zero return, we would not expect to observe *ex ante* negative real rates. Benjamin Friedman (1980) emphasizes the limited nature of portfolio substitution possibilities by arguing that if holding goods was a feasible portfolio alternative, then the expected rate of inflation could be viewed as simply the return from holding such real assets. In such a case, higher expected inflation would reduce the lenders supply of loans directly via portfolio substitutions from loans (bonds) to real goods. The problem with this scenario, as stated by Friedman is: "In practice, however, the available opportunities for such portfolio substitutions involving consumption or other real commodities are usually extremely limited; purchasing the Consumer Price Index basket of goods is not a feasible portfolio alternative" (p. 34).

The lack of such a real asset allows market instruments to offer a negative yield. Of course, with no inflation (and costless storage), one could hold money so that we don't expect to observe negative rates in noninflationary periods. Frederic Mishkin provides corroborating evidence in this area as he finds "The real rate, *whether adjusted or unadjusted for taxes*, is negatively correlated with inflation" (1981, p. 191, emphasis added) and "The real interest rate appears to have been positive in the 1950s and 1960s but has since turned negative in the mid- and late-1970s" (p. 192).

Regarding Tanzi's simulations, I do not believe that they correctly assess the path of the real rate following an exogenous change in inflationary expectations ($\pi$). A major theme of my earlier paper was the emphasis

---

*Arizona State University. I thank Don Schlagenhauf, John Schroeter, and Richard L. Smith for helpful discussions.

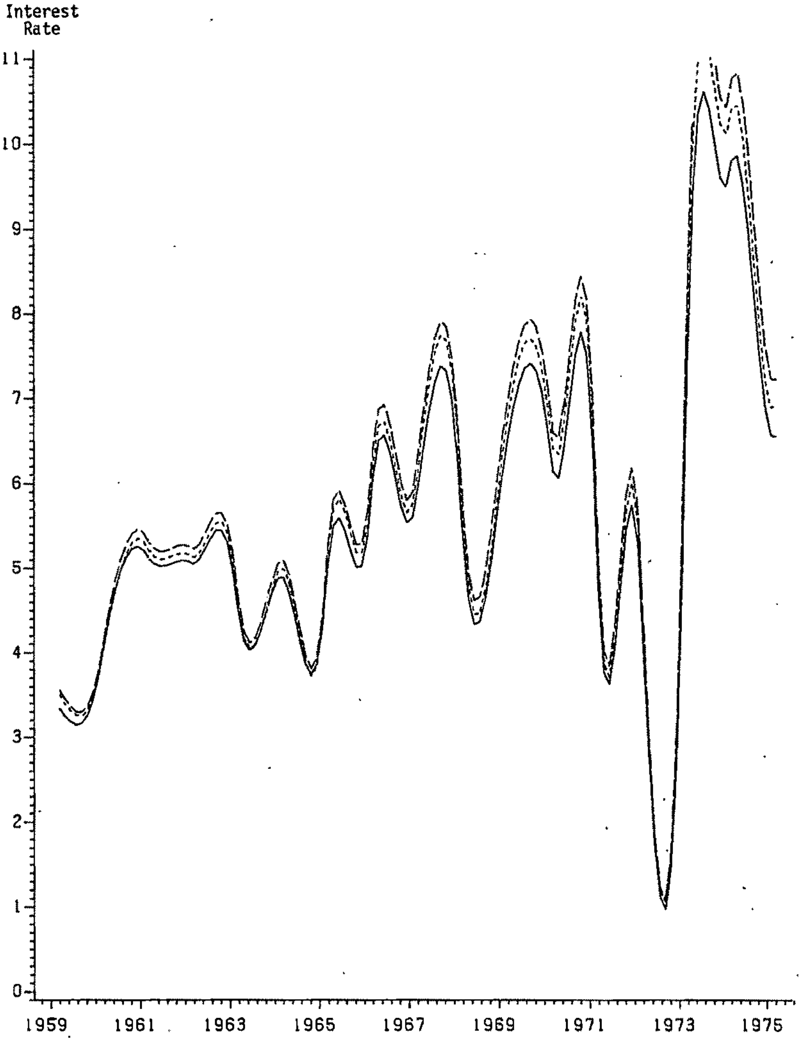


FIGURE 1. PRE-TAX NOMINAL INTEREST RATE SIMULATIONS
LEGEND: ———— BASE SIMULATION;
– – – – INCREASE EXPECTED INFLATION 10 PERCENT;
—— —— —— BOTH EXPECTED INFLATION AND INFLATION UNCERTAINTY INCREASED 10 PERCENT

placed on evaluating the Fisher equation and Darby hypothesis in the context of a reasonably comprehensive macro model. To illustrate the differences between a partial derivative approach and a total derivative, I have dynamically simulated the model presented in my earlier paper.[1] Figure 1 presents the solved values of the pre-tax nominal interest rate from three alternative simulations.

---

[1] In addition to Tanzi's failure to allow for the indirect effects of inflation on interest that are present in the structural model simulation, his simulations may also be criticized on the basis of the range of values allowed for expected inflation. While Tanzi considers $\pi$ from 2 to 20 percent over the 1959–75 period studied by Tanzi and myself, the mean value of $\pi$ was 2.73 percent and in all cases fell between .19 and 7.67 percent. If we allow our simulation experiment to wander far from the actual experience, then the results should be considered with great caution. In the case of inflation and interest, there is considerable evidence that the model parameters will change over time with an underlying change in the monetary experience (see my forthcoming paper, or Alexander Holmes and Myron Kwast, 1979), so that simulations incorporating inflation expectations that differ greatly from those used for model estimation may be meaningless.

The solid line in the figure gives the path of the base simulation using the data used to estimate the model coefficients in my 1982 article. The short-dashed line gives $i$ values when the expected rate of inflation is 10 percent higher than the control values (or $1.1\pi$). In my earlier study, the estimated coefficient on $\pi$ in the reduced-form nominal interest rate equation was .55. Yet the simulation results indicate that the total effect of a change in $\pi$ on $i$ is greater than .55. An increase in expected inflation of 10 percent, results in a simulated value of $i$ uniformly greater than that of the base simulation. Evaluating at the mean of each series, the results of these first two simulations, suggest a total derivative, $di/d\pi$, equal to .73.

While the total effect, .73, is certainly greater than the partial effect, .55, indicated by the reduced-form coefficient on $\pi$, there is reason to believe that a simulation based solely on increasing $\pi$ by 10 percent is quite unrealistic. The most obvious shortcoming is to assume that inflation uncertainty ($\sigma$), an exogenous variable proxied by the standard deviation of the price forecasts determining $\pi$, is invariant to changes in $\pi$, the mean expected inflation rate. Over the 1959–75 period studied, the correlation coefficient between $\pi$ and $\sigma$ was .90. Given that it is extremely unlikely that $\pi$ will change while $\sigma$ is constant, the simulation was reestimated with both $\sigma$ and $\pi$ increased by 10 percent. The longer-dashed line of Figure 1 lists the solved values of $i$. Comparing the new simulation to the base run and evaluating at the mean, we now find the total derivative $di/d\pi$ equal to 1.30.[2] The simulations confirm the point made in my earlier paper that while the reduced-form parameter on expected inflation is less than unity, this in no way invalidates the Darby hypothesis that the nominal interest rate will rise by more than the expected rate of inflation. Also, the simulations presented here serve to point out the limitations of Tanzi's partial derivative-based simulations.

While consideration of both indirect and direct effects of expected inflation on interest rates has suggested that the pre-tax nominal rate rises by more than the rise in expected inflation, it must be noted that the rise is not sufficient to prevent the simulated value of the after-tax real rate from turning negative by 1972. So while I think Tanzi's simulations were rather ill-conceived, I ultimately return to the basic problem he identifies: a negative after-tax real rate.

Issues involving unobservable phenomena like fiscal illusion are difficult to resolve. While the simulations presented above have hopefully offered a better insight into the interest rate effects of expected inflation than Tanzi's simulations, such exercises cannot settle the issue. Since the notion of financial markets ignoring tax effects is very improbable, it seems that those who argue for the existence of fiscal illusion have a responsibility to provide a theoretical basis for such ignorance. Considering the corporate and personal resources devoted to tax accountants and lawyers, and "tax-advantaged" investments, not to mention the relation between tax-free municipal bond interest rates and taxable bond rates, there would appear to be a preponderance of evidence suggesting that investors are quite cognizant of the effects of taxes on the real value of interest payments and receipts.

Given the incomplete state of markets and the consequent lack of an asset guaranteeing at least a zero net return during an inflation, the presence of negative after-tax real rates is not proof of fiscal illusion.

## REFERENCES

**Friedman, Benjamin M.,** "Price Inflation, Portfolio Choice, and Nominal Interest Rates," *American Economic Review*, March 1980, *70*, 32–48.

**Holmes, Alexander B. and Kwast, Myron L.,** "In-

---

[2] The notion of a total derivative assumes that $\pi$ affects $i$ through both direct and indirect channels. One such indirect channel is via $\sigma$. While $\pi$ and $\sigma$ are both given to the model, the spirit of this discussion implies that as agents reformulate expectations regarding the mean expected inflation rate, they also adjust the confidence interval about this mean. Experience has taught that a higher mean expected inflation rate leads to a greater variance about this mean. Ultimately, of course, both $\pi$ and $\sigma$ are determined by exogenous events like policy announcements. Practical considerations force us to make assumptions like $\sigma$ is a function of $\pi$.

terest Rates and Inflationary Expectations: Tests for Structural Change 1952–1976," *Journal of Finance*, June 1979, *34*, 733–41.

**Melvin Michael,** "Expected Inflation, Taxation and Interest Rates: The Delusion of Fiscal Illusion," *American Economic Review*, September 1982, *72*, 841–45.

______, "The Vanishing Liquidity Effect of Money on Interest: Analysis and Implications for Policy," *Economic Inquiry*, forthcoming.

**Mishkin, Frederic S.,** "The Real Interest Rate: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy: The Costs and Consequences of Inflation*, Autumn 1981, *15*, 151–200.

**Tanzi, Vito,** "Inflationary Expectations, Economic Activity, Taxes and Interest Rates," *American Economic Review*, March 1980, *70*, 12–21.

______, "Expected Inflation and Interest Rates: Comment," *American Economic Review*, June 1983, *73*, 501–02.

# Auditors' Report

February 23, 1983

Executive Committee
The American Economic Association

We have examined the balance sheets of The American Economic Association as of December 31, 1982 and 1981 and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of the American Economic Association as of December 31, 1982 and 1981, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Touche Ross and Co.
Certified Public Accountants
Nashville, Tennessee

THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS, DECEMBER 31, 1982 AND 1981

| | 1982 | 1981 |
|---|---|---|
| **Assets** | | |
| CASH | $ 640,267 | $ 557,710 |
| INVESTMENTS, at market (Notes A and B) | 2,808,085 | 1,911,347 |
| ACCOUNTS RECEIVABLE, less allowance for doubtful accounts of $337 (1982) and $1,243 (1981) | 76,375 | 81,595 |
| INVENTORY OF *Index of Economic Articles*, at cost | 37,511 | 65,631 |
| PREPAID EXPENSES | 15,022 | 15,065 |
| OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $17,622 (1982) and $14,371 (1981) | 27,276 | 23,284 |
| | $3,604,536 | $2,654,632 |
| **Liabilities and Fund Balances** | | |
| ACCOUNTS PAYABLE AND ACCRUED LIABILITIES | $ 361,952 | $ 285,883 |
| DEFERRED REVENUE (Note A): | | |
| Life membership dues | 52,302 | 54,924 |
| Other membership dues | 479,474 | 445,862 |
| Subscriptions | 415,751 | 411,037 |
| *Job Openings for Economists* | 16,986 | 15,555 |
| | 964,513 | 927,378 |
| ACCRUAL FOR DIRECTORY (Note A) | 87,141 | 33,318 |
| FUND BALANCES: | | |
| Restricted | 42,636 | 17,731 |
| General | 2,110,632 | 1,569,287 |
| Unrecognized change in market value of investments (Notes A and C) | 37,662 | (178,965) |
| Net Worth | 2,148,294 | 1,390,322 |
| Total Fund Balances | 2,190,930 | 1,408,053 |
| | $3,604,536 | $2,654,632 |

See notes to financial statements.

## The American Economic Association Statements of Revenues and Expenses for the Years Ended December 31, 1982 and 1981

| | 1982 | 1981 |
|---|---|---|
| Revenues from Dues and Activities: | | |
| Membership dues and subscriptions | $ 729,492 | $ 682,186 |
| Nonmember subscriptions | 645,362 | 620,681 |
| *Job Openings for Economists* subscriptions | 24,980 | 21,678 |
| Advertising | 87,350 | 98,946 |
| Sale of *Index of Economic Articles* | 64,171 | 65,752 |
| Sale of copies, republications, and handbooks | 31,813 | 33,558 |
| Sale of mailing list | 38,215 | 38,540 |
| Annual meeting | 36,854 | 14,491 |
| Sundry | 41,428 | 50,578 |
| | 1,699,665 | 1,626,410 |
| Investment Gains (Note B) | 329,584 | 110,224 |
| **Net Revenues** | **2,029,249** | **1,736,634** |
| Publication Expenses: | | |
| *American Economic Review* | 451,591 | 407,226 |
| *Journal of Economic Literature* | 626,065 | 564,311 |
| Directory publication (Note A) | 55,000 | 65,000 |
| *Job Openings for Economists* | 46,029 | 41,575 |
| *Index of Economic Articles* | 28,712 | 26,835 |
| | 1,207,397 | 1,104,947 |
| Operating and Administrative Expenses: | | |
| General and administrative: | | |
| Salaries | 141,980 | 124,725 |
| Rent | 12,016 | 11,232 |
| Other (Exhibit I) | 156,641 | 117,081 |
| Committee | 49,304 | 39,199 |
| Annual meeting | 5,400 | 3,640 |
| Provision for federal income taxes (Note A) | 2,500 | 5,000 |
| | 367,841 | 300,877 |
| **Total Expenses** | **1,575,238** | **1,405,824** |
| Revenues in Excess of Expenses | $ **454,011** | $ **330,810** |

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

| | Total | Operations | Market Value Adjustments |
|---|---|---|---|
| **Balance at December 31, 1980** | **$1,083,095** | **$590,806** | **$492,289** |
| Add market value adjustments resulting from inflation (Note A) | 155,382 | – | 155,382 |
| Add revenues in excess of expenses | 330,810 | 330,810 | – |
| **Balance at December 31, 1981** | **1,569,287** | **921,616** | **647,671** |
| Add market value adjustments resulting from inflation (Note A) | 87,334 | – | 87,334 |
| Add revenues in excess of expenses | 454,011 | 454,011 | – |
| **Balance at December 31, 1982** | **$2,110,632** | **$1,375,627** | **$735,005** |

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

| | Balance at January 1 | Receipts | Disbursements | Balance at December 31 |
|---|---|---|---|---|
| YEAR ENDED DECEMBER 31, 1981: | | | | |
| The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics | $ – | $ 125,727 | $ 115,227 | $ 10,500 |
| The Minority scholarship fund for minority students applying for graduate work in economics | 5,000 | – | – | 5,000 |
| Sundry | 1,111 | 3,626 | 2,506 | 2,231 |
| | **$6,111** | **$129,353** | **$117,733** | **$17,731** |
| YEAR ENDED DECEMBER 31, 1982: | | | | |
| The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics | $ 10,500 | $114,500 | $123,090 | $1,910 |
| The Minority scholarship fund for minority students applying for graduate work in economics | 5,000 | – | – | 5,000 |
| The Rockefeller Foundation grant for minority students applying for graduate work in economics | $ – | 58,500 | 25,105 | 33,395 |
| Sundry | 2,231 | 100 | – | 2,331 |
| | **$17,731** | **$173,100** | **$148,195** | **$42,636** |

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION FOR THE YEARS ENDED DECEMBER 31, 1982 AND 1981

| | 1982 | 1981 |
|---|---|---|
| **Cash**, beginning of year | $557,710 | $393,920 |
| SOURCES OF CASH: | | |
| Revenues in excess of expenses | 454,011 | 330,810 |
| Noncash charges: | | |
| Depreciation | 2,609 | 2,077 |
| Directory publication (Note A) | 55,000 | 65,000 |
| Market value adjustments (Note A) | (120,654) | 48,353 |
| Cash provided by operations | 390,966 | 446,240 |
| INCREASE (DECREASE) IN CASH DUE TO CHANGES IN: | | |
| Investments | (896,738) | 13,533 |
| Accounts receivable | 5,220 | 13,777 |
| Inventory of *Index of Economic Articles* | 28,120 | (10,074) |
| Prepaid expenses | 43 | 6,949 |
| Office furniture and equipment | (6,601) | (1,421) |
| Accounts payable and accrued liabilities | 76,069 | 71,810 |
| Deferred revenue | 37,135 | 78,147 |
| Accrual for directory | (1,177) | (170,784) |
| Restricted funds | 24,905 | 11,620 |
| General fund, market value adjustment | 87,334 | 155,382 |
| Unrecognized change in market value of investments | 337,281 | (451,389) |
| **Cash**, end of year | $ 640,267 | $ 557,710 |

See notes to financial statements.

## Notes to Financial Statements

### A. Summary of Significant Accounting Policies

*Investments* are accounted for on a market value basis. According to the method the Association uses to value investments, the change in market value of corporate stocks, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (4.6% in 1982 and 8.6% in 1981), is recognized in income over a three-year period for corporate stocks and reflected in current income for government obligations, bonds and commercial paper.

*The accrual for directory* results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1981 and distributed at no cost to the membership. In order to match more properly the publishing cost of this directory with revenue from membership dues, the Association provided $55,000 in 1982 and $65,000 in 1981 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

*Deferred revenue* represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

*The American Economic Association* files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

**B. Investments and Investment Income**

Investments consist of:

| | December 31, 1982 | | December 31, 1981 | |
|---|---|---|---|---|
| | Cost | Market | Cost | Market |
| Government obligations, bonds, and commercial paper | $ 829,121 | $ 927,110 | $ 514,807 | $ 516,537 |
| Corporate stocks | 1,237,507 | 1,880,975 | 1,133,260 | 1,394,810 |
| | $2,066,628 | $2,808,085 | $1,648,067 | $1,911,347 |

Investment gains (losses) recognized consist of:

| | Year Ended December 31 | |
|---|---|---|
| | 1982 | 1981 |
| Government obligations, bonds, and commercial paper: | | |
| Interest | $146,549 | $103,813 |
| Increase (decrease) in market value recognized | 72,091 | (33,995) |
| | 218,640 | 69,818 |
| Corporate stocks: | | |
| Cash dividends | 62,381 | 54,764 |
| Increase (decrease) in market value recognized (Note C) | 48,563 | (14,358) |
| | 110,944 | 40,406 |
| Investment gains, net | $329,584 | $110,224 |

**C. Unrecognized Change in Market Value of Investments**

As described more fully in Note A, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1982 and 1981 revenues, and the amount of these market value increases (decreases) that will be recognized in income in future periods.

| Year of Market Value Change | Recognized in Income in | | To be Recognized in | | Unrecognized Change | |
|---|---|---|---|---|---|---|
| | 1982 | 1981 | 1983 | 1984 | 1982 | 1981 |
| 1979 | $ – | $ 25,475 | $ – | $ – | $ – | $ – |
| 1980 | 99,298 | 99,298 | – | – | – | 99,298 |
| 1981 | (139,131) | (139,131) | (139,131) | – | (139,131) | (278,263) |
| 1982 | 88,396 | – | 88,396 | 88,397 | 176,793 | – |
| | $ 48,563 | ($ 14,358) | $ 50,735 | ($ 88,397) | ($ 37,662) | ($178,965) |

**D. Retirement Annuity Plan**

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately $21,000 and $18,000 for 1982 and 1981, respectively.

### EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1982 AND 1981

| | 1982 | 1981 |
|---|---|---|
| Dues and subscriptions | $ 42,405 | $ 9,400 |
| Mailing list file maintenance | 20,535 | 22,455 |
| Postage | 19,751 | 18,377 |
| Periodic mailing expenses | 14,019 | 13,017 |
| Accounting and legal | 12,200 | 12,150 |
| Office supplies | 11,271 | 14,520 |
| Economic Institute Fellowships | 10,000 | - |
| Investment counsel and custodian fees | 8,474 | 7,242 |
| President and president-elect expenses | 4,368 | 4,900 |
| Insurance and miscellaneous | 4,062 | 5,739 |
| Telephone | 4,010 | 4,074 |
| Depreciation (straight-line method) | 2,609 | 2,077 |
| Currency exchange charges | 1,614 | 1,248 |
| Travel and entertainment | 783 | 1,370 |
| Uncollectible receivables | 540 | 512 |
| | **$156,641** | **$117,081** |

# NOTES

### *1984 Nominating Committe of AEA*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Charles L. Schultz has appointed a Nominating Committee for 1984 consisting of William J. Baumol, Chair; Clair Brown, Robert M. Coen, John J. Gurley, Stanley Lebergott, Glenn C. Loury, and Richard H. Thayler. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

### *Nominations for AEA Officers: 1984*

The Electoral College on March 18 chose Charles P. Kindleberger as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1983. Other nominees (chosen by the 1983 Nominating Committee) are: Vice President (two to be elected), Hollis B. Chenery, Zvi Griliches, Robert L. Heilbroner, and Henry J. Aaron; for members of the Executive Committee (two to be elected), Michael R. Darby, Stanley Lebergott, Janet L. Norwood, and Victor R. Fuchs.

Under a change in the bylaws as described in the *American Economic Review Proceedings*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

---

*New Hertz Car Rental Rate*: Effective February 1, 1983, AEA members are entitled to a 15 percent discount on "Standard Unlimited Mileage Rates" at all Hertz corporate and participating licensee locations in the United States. This discount applies to basic or standard rates in the United States only. It does not apply to prediscounted rates such as touring rates, economy fares, weekend specials, etc. The old discount rate was 5 percent; see the Notes section of this *Review*, December 1982, p. 1214.

---

Proposals are invited by the Japan–U.S. Educational Commission from American and Japanese scholars and professionals for a special project on U.S.–Japan Economic and Trade Relations. The proposal must include a detailed explanation of the research including its relevance to economic and trade relations between the United States and Japan; curriculum vitae; a projected schedule not to exceed two years; and a budget. The size of the grants will depend on the nature of the submitted proposals. The maximum grant would be approximately $100,000. The application deadline is September 30, 1983. For further information or forms, contact Jennifer Keefe, Program Officer, Council for International Exchange of Scholars, 11 Dupont Circle, Suite 300, Washington, D.C. 20036. (Telephone 202+833-4981.) In Japan, contact the Japan–U.S. Educational Commission, 2-14-2, Nagata-cho, Chiyoda-ku, Japan 100.

---

The Institute for Research on Poverty at the University of Wisconsin-Madison, in association with the Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, is sponsoring a Small Grants–Visitors Program for social scientists doing empirical poverty research related to a variety of topics, including trends in economic status and the efficiency and equity effects of income transfer policies. The Program will be from July 1, 1983 to December 31, 1984. For further information, contact the Director, Institute for Research on Poverty, 3412 Social Science Building, University of Wisconsin, Madison, WI 53706.

---

The Association for Canadian Studies in the United States will hold its seventh biennial conference September 29–October 1, 1983, at the Samoset Resort in Rockport, Maine. All areas of Canadian studies will be represented. For further details, contact the ACSUS National Office, 1776 Massachusetts Ave., NW, Washington, D.C. 20036. (Telephone 202+822-8688.)

---

Nominations for candidates for Sloan Research Fellowships are being accepted for the period beginning September 1984. The deadline is September 15, 1983. Candidates must be members of the faculty at a college or university in the United States or Canada, and must be at an early stage of their research careers. For information and nomination forms, contact Fellowships for Basic Research, Alfred P. Sloan Foundation, 630 Fifth Avenue, New York, NY 10111.

---

The ninth annual meeting for the Society for Historians of American Foreign Relations will be held at the Catholic University of America, Washington, D.C., August 4–6, 1983. Further information may be obtained from Alan K. Henrikson, Chairman, SHAFR Program Committee, The Fletcher School of Law and Diplomacy, Tufts University, Medford, MA 02155.

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics announce the institution of an annual Leonard J. Savage Award of five hundred dollars ($500) for an outstanding doctoral dissertation in the area of Bayesian Econometrics and Statistics. To be considered for the 1983 Savage Award, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1983, accompanied by a short letter from the supervisor summarizing the main results of the dissertation. Dissertations completed after January 1, 1977 are eligible to be considered for the 1983 Savage Award. An Evaluation Committee will be appointed by the board of the Leonard J. Savage Memorial Trust Fund (M. H. DeGroot, S. E. Fienberg, S. Geisser, J. B. Kadane, E. E. Leamer, J. W. Pratt, and A. Zellner, Chairman) to evaluate dissertations that are submitted for the Savage Award. Dissertations and supporting letters should be sent to Professor Arnold Zellner, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637.

The winner of the 1982 Leonard J. Savage Thesis Award is Soo Hong Chew, "Two Representation Theorems and Their Application to Decision Theory," completed at the University of British Columbia. The Thesis Evaluation Committee included Bruce M. Hill (Chairman), Nicholas M. Kiefer, David Lane, and Arnold Zellner (*ex officio*).

The fifteenth annual meeting of the International Society for the History of Behavioral and Social Sciences (CHEIRON), will be held June 15–18, 1983, at Glendon College, York University, Toronto, Ontario, M4N 3M6.

Global Risk Assessments, Inc. is publishing *Global Risk Assessments*: *Models and Methodologies*, Volume II, scheduled for release during winter 1984. Articles and monographs dealing with issues of political risk assessment and management are requested. Especially desired are articles relating to actual corporate experiences in this area, formatted as case studies showing application of models and methods. All submissions will be reviewed by an Editorial Board. The deadline is October 31, 1983. Please forward all manuscripts to Dr. Jerry Rogers, President; Global Risk Assessments, Inc., 3638 University Avenue, Suite 215, Riverside, CA 92501.

The School of Urban and Public Affairs at Carnegie-Mellon University, with the support of a training grant from the Center for Studies in Crime and Delinquency of the National Institute of Mental Health, is offering a postdoctoral program in Quantitative Methods in Criminal Justice. In addition to a stipend of at least $13,380 (depending on years of postdoctoral experience), all training costs and research resources are provided by the training grant. A limited number of predoctoral fellowships are also available to individuals who have already completed at least two years of graduate study. Participation in the program can begin in July 1983. Applications should be submitted as early as possible. For further information and for application forms, write to Professor Alfred Blumstein, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA 15213.

The Council for International Exchange of Scholars announces the 1984–85 competition for Senior Scholar Fulbright awards for university teaching and postdoctoral research. Awards are offered in all academic fields for periods of 2–10 months in over 100 countries. Prospective applicants may write for applications and additional details on awards, specifying the country and field of interest, to: Council for International Exchange of Scholars, 11 Dupont Circle, Suite 300, Washington, D.C. 20036. All applicants must be U.S. citizens and have had college or university teaching experience and/or a Ph.D. or the equivalent. The applications deadlines are June 15, 1983, for American Republics, Australia, and New Zealand, and September 15, 1983, for Africa, Asia, Europe, and the Middle East.

Twelve long-term (6–10 months) and nine short-term (2–3 months) research awards, without restriction as to field, are offered for 1984–85 by the Indo-U.S. Subcommission on Education and Culture. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. Fellowship terms include: $1,200–$1,500 per month, depending on academic/professional achievement and seniority, $350 per month payable in dollars and the balance in rupees; an allowance for books and study/travel in India; and international travel for the grantee. In addition, long-term fellows receive international travel for dependents; a dependent allowance of $100–$250 per month in rupees; and a supplementary research allowance up to 34,000 rupees. This program is sponsored by the Indo-U.S. Subcommission on Education and Culture and is funded by the U.S. Information Agency, the National Science Foundation, the Smithsonian Institution, and the Government of India. *The application deadline is July 1, 1983.* Application forms and further information are available from the Council for International Exchange of Scholars, Attention: Indo-American Fellowship Program, Eleven Dupont Circle, Suite 300, Washington, D.C. 20036. (Telephone 202 + 833-4985.)

*Call for Papers*: The annual meeting of the Gesellschaft für Wirtschafts- und Sozialwissenschaften- Verein für Socialpolitik, "World Economic Development: Problems and Perspectives," will be held September 17–19, 1984. Themes will include Trade and Trade Policy: Primary Sector and Secondary and Tertiary Sector; International Capital Transactions and Balance of Payments; the Process of Innovation and Labor Migration; Institutions; and Econometric World Models. Send 3–5 page abstracts, a brief curriculum vitae, and a list of publications (in triplicate) by August 1, 1983, to the Chairman of the Preparatory Committee: Professor Herbert Giersch, Kiel Institute of World Economics, P.O. Box 43 09, 2300 Kiel 1, Federal Republic of Germany. (Telephone 0431/884-236.)

The International Society for Inventory Research (ISIR) was established at the August 1982 meeting of the Second International Symposium on Inventories. The ISIR provides those engaged in inventory research with an opportunity to exchange views and report on research progress. The constitution for the society provides for two official languages, English and Russian. The next meeting is tentatively scheduled for August 1984 in Budapest. Further information can be obtained from the President, Professor Kenneth J. Arrow, Department of Economics, Stanford University, Stanford, CA 94305, or the Secretary General, Dr. Attila Chikan, Karl Marx University of Economics, Budapest Pf. 489, Hungary, H-1882.

The John M. Olin Fellowship Program in Law and Economics, administered by the Law and Economics Center, Emory University, is accepting applications for the class entering in August 1984. The three-year program leads to the Doctor of Law degree and is designed for individuals with a strong foundation in graduate microeconomics and its application. Five fellowships are available each year that provide full tuition and fees (currently about $6,200 per year) plus stipends of at least $8,000 per year ($7,000 per year for Fellows without the Ph.D.). Up to five additional fellowships of lesser amounts may also be offered each year. The deadline for submitting applications is March 1, 1984. Candidates must apply separately to the School of Law and should register for the Law School Admissions Test no later than January 1984. All awards will be made no later than April 1, 1984. Early applicants may request early decisions. For information, write to the John M. Olin Fellowship Program, Law and Economics Center, Emory University, Altanta, GA 30322. Members of the Center will attend the meetings of the WEA (Seattle, July 20–24), the SEA (Washington, November 20–22), and the AEA (San Francisco, December 28–30). Interested persons should request information and application packets, and may reserve time for an interview at any convenient meeting listed.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

## Deaths

Leif Johansen, professor, Institute of Economics, University of Oslo, December 29, 1982.

Benjamin Joseph Katz, professor of economics, New York University, October 1, 1982.

Fritz Machlup, professor of economics at New York University and professor emeritus at Princeton University, January 30, 1983.

Wesley S. Mellow, Office of Research and Evaluation, U.S. Bureau of Labor Statistics, December 10, 1982.

## Retirement

Charles J. Kennedy, professor of economics and business history, University of Nebraska-Lincoln, May 1982.

John F. Timmons, professor of economics, Iowa State University, December 31, 1982.

## Promotions

Sham L. Bhatia: associate professor of economics, Indiana University-Northwest, July 1, 1982.

Richard V. Burkhauser: associate professor of economics, Vanderbilt University, September 1, 1982.

Christine M. Cumming: chief, Industrial Economies Division, international research department, Federal Reserve Bank of New York, January 1, 1983.

James L. Dietz: professor of economics, California State University-Fullerton, September 1982.

Peter Fousek: executive vice president and director of research, Research and Statistics Function, Federal Reserve Bank of New York, January 1, 1983.

Richard J. Gelson: vice president, Research and Statistics Function, Federal Reserve Bank of New York, January 1, 1983.

Laurie S. Goodman: senior economist, financial markets department, Federal Reserve Bank of New York, January 13, 1983.

Roger M. Kubarych: senior vice president and deputy director, Research and Statistics Function, Federal Reserve Bank of New York, January 1, 1983.

Arturo C. Porzecanski: vice president, International Economics Department, Morgan Guaranty Trust Co. of New York, November 1, 1982.

Irwin D. Sandberg: senior vice president, Foreign Relations Function, Federal Reserve Bank of New York, January 1, 1983.

Vincenzina Santoro: vice president, Morgan Guaranty Trust Co., November 1982.

Richard C. Schiming: associate professor of economics, Mankato State University, September 15, 1982.

Eric J. Solberg: professor of economics, California State University-Fullerton, September 1982.

George H. Sweeney: associate professor of economics, Vanderbilt University, September 1, 1982.

Miguel A. Valdivieso: manager, financial planning and economics, Regional Transportation Authority, October 1982.

John A. Wenninger: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, January 1, 1983.

### Administrative Appointments

George Blyn: chairman, department of economics, Camden College, Rutgers University, January 1, 1983.

Richard V. Butler, Temple University: chairman, department of economics, Trinity University, July 1982.

Joseph Horton: dean, W. Fielding Rubel School of Business, Bellarmine College, Fall 1982.

Arthur Kraft, University of Nebraska-Lincoln: dean, College of Business and Economics, West Virginia University, July 1, 1983.

Eric J. Solberg: chair, department of economics, California State University-Fullerton, September 1982.

William B. Walstad: director of economic education, University of Nebraska-Lincoln, August 1982.

### Appointments

Moshe Adler: assistant professor of economics, University of California-Davis, July 1, 1982.

Juan Amieva-Huerta: assistant professor economics, Anahuac University, Mexico, Spring 1983.

Bruce R. Bartlett: executive director, Joint Economic Committee, U.S. Congress, January 3, 1983.

John Bender, University of Chicago: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, October 1982.

David M. Blau, Yale University: assistant professor of economics, University of Miami, August 1982.

William L. Breit, University of Virginia: professor of economics, Trinity University, January 1983.

Richard V. Butler, Temple University: associate professor of economics, Trinity University, July 1982.

Peter J. Coughlin, Carnegie-Mellon University: assistant professor of economics, University of Maryland-College Park, August 1982.

Mark Dynarski: assistant professor of economics, University of California-Davis, July 1, 1981.

Elizabeth Goldstein: economist, Developing Economies Division, external financing department, Federal Reserve Bank of New York, November 1, 1982.

William Hall, University of Michigan: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, September 1982.

Timothy Hau: assistant professor of economics, University of California-Davis, July 1, 1981.

Kenneth Heyer, University of California-Los Angeles: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, October 1982.

Barbara H. Kehrer, Mathematica Policy Research: senior program officer, The Robert Wood Johnson Foundation, November 15, 1982.

Kenneth Kletzer: assistant professor of economics, University of California-Davis, July 1, 1982.

Kishore G. Kulkarni, University of Pittsburgh: assistant professor of economics and finance, University of Central Arkansas, August 15, 1982.

Douglas O. Love: adjunct assistant professor of economics, University of Nebraska-Lincoln, August 1982.

LeAnn McGranahan: adjunct assistant professor of economics, Iowa State University.

Frank J. P. Pinto: secretary, United Nations Interagency Technical Energy Group, and UNDP Technical Energy Coordinator with the UN Department of International Economic and Social Affairs, New York, October 1981.

Philip K. Robins, SRI International: professor of economics, University of Miami, August 1982.

Daniel H. Saks, Brookings Institution: professor of educational policy, George Peabody College for Teachers, and professor of economics, Vanderbilt University, September 1983.

Theodore R. Valentine, Vanderbilt University: visiting assistant professor of economics, Trinity University, July 1982.

William B. Walstad: associate professor of economics, University of Nebraska-Lincoln, August 1982.

David Weisberg, Stanford University: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, January 1983.

Michael Williams, University of Chicago: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, November 1982.

### Leaves for Special Appointments

LeAnn McGranahan, Iowa State University: Agency for International Development, Zambia.

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

---

## NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1983 issue of the *Review* will carry the eightieth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1983. This announcement is an invitation to send us information for the preparation of the list.

The *Review* will publish in its December 1983 issue the names of those who will have been awarded the doctoral degree since June 1982 and the titles of their dissertations. Dissertation abstracts are not published, as these are published elsewhere.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

*JEL* Classification No. ________

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ____________________

Degree Conferred (Ph.D. or D.B.A.) __________ Year ____________

Dissertation Title: ____________________

## From our first—

## JOURNAL OF POLITICAL ECONOMY

The best general interest economics journal (according to a survey of the AEA), **JPE** publishes analytical, interpretive, and empirical studies in all traditional areas of economics (including monetary theory, fiscal policy, labor economics, planning and development, international trade, and finance) as well as interdisciplinary studies. *Published bimonthly.* **Edited by Jacob A. Frenkel, James J. Heckman, Sam Peltzman, and George J. Stigler.**

## —to our latest

## JOURNAL OF LABOR ECONOMICS

Sponsored by the Economics Research Center/NORC, this new journal fills the urgent need for a high-quality international periodical that presents both theoretical and applied papers in the rapidly expanding field of labor economics. Topics covered include supply and demand for labor services, compensation, labor markets, distribution of income, labor demographics, unions and collective bargaining, and applied and policy issues in labor economics. *Published quarterly.* **Edited by Edward P. Lazear. Associate Editors: Glenn M. T. MacDonald, George R. Neumann, and Robert Topel.**

**Among the contributors to JOLE, volume one:**

*Lorne Carmichael*
*Barry R. Chiswick*
*Christopher J. Flinn and James J. Heckman*
*W. Gould, W. McManus, and F. Welch*
*Giora Hanoch and Marjorie Honig*
*H. Gregg Lewis*
*Robert H. Meyer and David A. Wise*
*John Raisian*
*Chitra Ramaswami*
*Sherwin Rosen*
*Burton A. Weisbrod*

## You've come to expect quality economics journals

**JPE/JOLE** ☐ **JOLE** charter 1-year rate (to begin with vol. 1, no. 1, January 1983):
☐ Individuals $25 (reg. $30) ☐ Institutions $50 (reg. $60)
☐ Students (with copy of ID) $20 (reg. $24)
☐ I haven't yet decided. Please send me a sample copy of **JOLE.**

☐ **JPE** 1-year rates: ☐ Individuals $29 ☐ Institutions $50
☐ Students (with copy of ID) $18

Name ____________________
Address ____________________
City ____________ State/Country ____________ ZIP ________

**Visa and MasterCard accepted.** Please mail this coupon with payment, purchase order, or complete charge card information to The University of Chicago Press, P.O. Box 37005, Chicago, IL 60637. AEA 6/83 KO1AB

## from The University of Chicago Press

Times Mirror College Publishing
Proudly Announces The Appointment of

**Kenneth J. Arrow**

**as**

**Consulting Editor in Economics**

**Professor, Stanford University**
**1972 Nobel Memorial Prize**
**Past President of the American Economic Association**

Times Mirror College Publishing, a division of The C.V. Mosby Company, will begin publishing a full line of economics texts at both undergraduate and graduate levels. We are very pleased that Dr. Arrow will be advising us in this publishing effort.

If you would like to know more about our publishing plans, contact:
Alfred E. Hockwalt
Senior Acquisitions Editor
Times Mirror College Publishing
11830 Westline Industrial Drive
St. Louis, MO 63141
(800) 325-4177/(314) 872-8370

MOSBY
TIMES MIRROR

ASP188

Books that present incisive analyses of economic structure and policy...

## POVERTY AND POLICY IN AMERICAN HISTORY

MICHAEL B. KATZ

*A Volume in the STUDIES IN SOCIAL DISCONTINUITY Series*

Dr. Katz uses records of welfare clients, poorhouses, and studies of paupers, tramps, and other groups to demonstrate the integral role of dependence in working-class life. He illustrates the effects of the wide gap between beliefs about poor people and their actual experience and demography on social policy.

**1983, 304 pp., $27.50**
**ISBN: 0-12-401760-6 (cloth)**
**1983, 304 pp., $12.50**
**ISBN: 0-12-401762-2 (paper)**

## WORKERS' EARNINGS AND CORPORATE ECONOMIC STRUCTURE

RANDY HODSON

*A Volume in the STUDIES IN POPULATION Series*

In this study, Hodson examines the role of industry- and firm-level corporate structure in determining employees' earnings. He emphasizes market, technical, and organizational characteristics of the workplace and argues that corporate structures should be reconceptualized in terms of resource provision to workers and owners.

**1983, c. 238 pp., in preparation**
**ISBN: 0-12-351780-x**

## THE EXPORT–IMPORT BANK

### AN ECONOMIC ANALYSIS

DAVID P. BARON

*A Volume in the ECONOMIC THEORY, ECONOMETRICS, AND MATHEMATICAL ECONOMICS Series*

This volume critically analyzes the export financing issue and the Eximbank's efforts to fulfill its congressional mandate. Baron offers recommendations concerning Eximbank administrative procedures and United States export financing objectives and proposes strategies for achieving these objectives.

**1983, 360 pp., $39.50**
**ISBN: 0-12-079080-7**

*Two Volumes in the QUANTITATIVE STUDIES IN SOCIAL RELATIONS Series . . .*

## THE STRUCTURAL DETERMINANTS OF UNEMPLOYMENT

### VULNERABILITY AND POWER IN MARKET RELATIONS

PAUL G. SCHERVISH

This work proposes and tests a theory on the sociology of unemployment in the United States from 1969 to 1978. The author defines the structural determinants of unemployment and examines three such determinants: class position, the product market characteristic of the firm, and the period of the business cycle. Economic outcomes are then linked to the effects of these determinants on the vulnerability and power of workers in the labor market.

**1983, 256 pp., $29.50**
**ISBN: 0-12-623950-9**

## ASCRIPTION AND LABOR MARKETS

### RACE AND SEX DIFFERENCES IN EARNINGS

TOBY L. PARCEL
and CHARLES W. MUELLER

The authors analyze earnings attainment as a function of personal and structural characteristics for workers who vary by race and sex. This investigation is set in the contexts of local, industrial, and occupational labor markets as well as measures of workers' class/authority positions.

**1983, 336 pp., $32.50**
**ISBN: 0-12-545020-6**

*Send payment with order and save postage and handling. Prices are in U.S. dollars and are subject to change without notice.*

**ACADEMIC PRESS, INC.**
*A Subsidiary of*
*Harcourt Brace Jovanovich, Publishers*
**New York • London • Toronto • Sydney**
**San Francisco**
**111 FIFTH AVENUE, N.Y., N.Y. 10003**

313323

# CAMBRIDGE

## Inside the Black Box

*Technology and Economics*

**Nathan Rosenberg**

Rosenberg breaks open the black box–the mythical container that economists have treated as the incubator for technological phenomena and have chosen not to peer into too closely. Examining the mechanism of technological change, he shows how specific features of individual technologies have shaped a number of variables of major concern to economists: the rate of productivity improvement; the nature of the learning process underlying technological change; the speed of technology transfer; and the effectiveness of government policies intended to influence technologies.

Hardcover $29.95 Paper $12.95

## What Political Economy Is All About

*An Exposition and Critique*

**Ezra Mishan**

Dr. Mishan discusses the basis for prescriptive economics and the rationale of the economist's efficiency criterion. This leads to considerations of standard allocation rules, the concept of consumer surplus and rent, and of second best and externalities. The book concludes with a discussion of sources of economic failure in a technological age.

Hardcover $32.50 Paper $9.95

*Now in paperback*

## The Years of High Theory

*Invention and Tradition in Economic Thought 1926–1939*

**G.L.S. Shackle**

A reissue, in paperback, of Professor Shackle's classic work first published in 1967, for which he has provided a new foreword. $14.95

## Structural Change and Economic Growth

*A Theoretical Essay on the Dynamics of the Wealth of Nations*

**Luigi L. Pasinetti**

Also in paperback, Pasinetti's original theoretical treatment of the problems of maintaining full employment.

"This is the culmination of Professor Pasinetti's work on growth and income distribution, to which he has already made such notable contributions. . . ."–*The Economic Journal* $14.95

## Exchange-Rate Determination

**Anne O. Krueger**

In this concise study, Professor Krueger analyzes the behavior of the international monetary system and economists' ideas about it in the post-war period. Contrasting the floating exchange rate system of the 1970s with a fixed exchange rate system, she presents a thorough survey of the modern theory of exchange rate determination.

*Cambridge Surveys of Economic Literature*

Hardcover $34.50 Paper $9.95

*Now in paperback*

## Asset Markets and Exchange Rates

*Modeling an Open Economy*

ABRIDGED EDITION

**Polly Reynolds Allen and Peter B. Kenan**

This paperback edition consists of the first three parts of *Asset Markets, Exchange Rates and Economic Integration*. These three parts stand alone, as the authors intended. Allen and Kenan synthesize and extend recent developments in international monetary theory using a general model of an open economy.

"An important achievement. It works out in detail and in a notably systematic fashion the implications of the new approach to balance of payments or exchange rate theory."
–*The Economic Journal* $17.95

## Limited-Dependent and Qualitative Variables in Econometrics

**G.S. Maddala**

Professor Maddala presents the econometric analysis of single equation models where the jointly dependent variables can be continuous, categorical, or truncated.

*Econometric Society Monographs in Quantitative Economics* $39.50

## International Industrial Productivity

*A Comparison of Britain, America, and Germany*

**A.D. Smith, D.M.W.N. Hitchens and S.W. Davies**

*NIESR Occasional Paper XXXIV* $24.95

### Rational Expectations

**Steven M. Sheffrin**

This book develops the logic behind the "rational expectations" approach, discussing in detail the progress that has been made in areas ranging from macroeconomics and financial markets to theoretical and applied microeconomics. The author also looks at how the current debate on such questions has shifted due to a perception of the shortcomings of these models and hypotheses.

*Cambridge Surveys of Economic Literature*

Hardcover $29.95 Paper $8.95

### Measuring Economic Welfare

*New Methods*

**George McKenzie**

Professor McKenzie presents a solution to one of the most important problems of applied economics: the formulation of operational procedures designed to make possible the evaluation of projects and economic policies. He discusses the new method based on Samuelson's "money-metric," uses examples to show its practicality, and compares it with alternative procedures. $32.50

### Investment in Education and Social Choice

**Tapas Majumdar**

An introduction to the new economics of education. Theoretical arguments are built around a primary distinction between individual decisions on investment in learning and institutional decisions on providing education. The resulting micro- and macroeconomic considerations and collective choice questions are then addressed. The author applies his arguments to examples from Brazil, India, France, Britain and the United States. $29.95

### Inflation Accounting

*An Introduction to the Debate*

**Geoffrey Whittington**

A concise, clear summary of the theory of inflation accounting that describes and illustrates all the major methods.

*Management and Industrial Relations 3*

Hardcover $29.95 Paper $9.95

### Labor Markets and Inequitable Growth

*The Case of Authoritarian Capitalism in Brazil*

**Samuel A. Morley**

The author challenges the view that Brazil's inequitable growth was caused by inequitable government policies and sees it rather as a result of skill-intensive growth in a labor surplus economy. He argues that the poor benefited more than the statistics imply, and compares Brazil with Taiwan and Korea, suggesting strategies that could have saved Brazil from uneven development. $39.50

### Brazil's State-Owned Enterprises

*A Case Study of the State as Entrepreneur*

**Thomas J. Trebat**

The first systematic analysis of the performance of Brazil's large state-owned enterprises. Trebat examines their economic and financial performance in terms of their contributions to economic growth and concludes that they have been effective substitutes for private investment in a number of strategic industries.

*Cambridge Latin American Studies 45*

about $39.50

### Public Enterprise in Less-Developed Countries

**Leroy P. Jones, Editor**

Seventeen papers selected from a Boston Area Public Enterprise Group Conference examine government-owned firms in less-developed countries that sell goods and services in a market. $39.50

### Central Grants to Local Governments

*The Political and Economic Impacts of the Rate Support Grant in England and Wales*

**R.J. Bennett**

*Cambridge Geographical Studies 17* $49.50

## Cambridge University Press

32 East 57th Street, New York, N.Y. 10022

# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA has joined a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by dividend credits. In the past seven years, insured members received credits on their April 1 semiannual payment notices averaging over 45% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future dividends and credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

**Administrator, AEA Group Insurance Program** E-2
**1707 L Street, N.W.—Suite 700**
**Washington, D.C. 20036**

Please send me more information about the AEA Life Insurance Plan.

Name ______________________ Age ______

Address ______________________

City ______________ State ______ Zip ______

Or—call today Toll-Free 800-424-9883
(Washington, DC area, call 296-8030)

New Books in

ECONOMICS

John M. Blatt

**DYNAMIC ECONOMIC SYSTEMS**

**A Post-Keynesian Approach**

"...one of the most intelligent contributions to, and certainly one of the clearest expositions of three major areas — growth theory, capital theory, and macroeconomics." — *G.C. Harcourt, Jesus College, Cambridge*

370 pages Cloth $35.00

Adolph Lowe

**ON ECONOMIC KNOWLEDGE**

**Toward a Science of Political Economics**

"This work is nothing less than a frontal challenge to the world view held by the economics profession." — *Journal of Economic Issues.*

*Expanded edition — for the first time in paperback.*

364 pages Paper $14.95

Peter E. Earl

**THE ECONOMIC IMAGINATION**

"...a beautifully written, sophisticated critique of orthodox economics and a genuine attempt to construct an alternative theory of choice....Essential reading for anyone having the slightest interest in the behavior of consumers and business firms." — *Mark Blaug, University of London Institute of Education*

224 pages Cloth $30.00

Malcolm C. Sawyer

**MACRO-ECONOMICS IN QUESTION**

**The Keynesian-Monetarist Orthodoxies and the Kaleckian Alternative**

"By his comparison of the microfoundations of [Monetarism and American Keynesianism] and those of the Kaleckian alternative, the student is introduced to a provocative and timely debate on the macro-economic issues of the era."
— *Leonard A. Rapping, University of Massachusetts, Amherst*

220 pages Cloth $25.00 Paper $14.95

*M.E. Sharpe, Inc.* **80 Business Park Drive**
**Armonk, New York 10504**

# LEADING ECONOMIC INDICATORS

## The Voice of the Poor
Essays in Economic and Political Persuasion
*John Kenneth Galbraith*

Galbraith reverses the usual order of wealthy nations dictating to poor and shows what the underdeveloped countries could teach the rich about the military, economic, and cultural relationships affecting both.
$8.95

## Two Hungry Giants
The United States and Japan in the Quest for Oil and Ores
*Raymond Vernon*

In this comparative analysis of policies and approaches, Vernon suggests what is likely to occur in the future as two economic superpowers search the world for raw materials.
$16.00

## Advanced Exercises in Microeconomics
*Paul Champsaur and Jean-Claude Milleron*
Translated by John P. Bonin and Hélène Bonin

Through a sequence of problems, this book gives a comprehensive view of resource allocation theory.
$27.50

## A Treatise on the Family
*Gary S. Becker*

"A major achievement. It is the most important book on the family to appear in many years."
—*Journal of Economic Literature*
$20.00

## How We Live
An Economic Perspective on Americans from Birth to Death
*Victor R. Fuchs*

"In pellucid prose, Victor Fuchs uses economics to enlarge our understanding of social problems, and adds greatly to our ability to think more carefully about a large range of policy issues. In a work studded with insights, he teaches us that to a considerable extent demography is destiny."—John Kaplan
$17.50

## Capital Taxation
*Martin Feldstein*

Feldstein shows how systems of taxation influence the rate and nature of capital formation. His identification of key economic and policy questions make this book a powerful addition to the economic literature.
$40.00

## An Evolutionary Theory of Economic Change
*Richard R. Nelson and Sidney G. Winter*

In the most sustained attack on mainstream neoclassical economics in more than forty years, Nelson and Winter borrow the concept of natural selection to construct an evolutionary theory of business behavior.
*Belknap* $25.00

**Risk by Choice**
Regulating Health and Safety in the Workplace
*W. Kip Viscusi*
Here is a comprehensive analysis of the economic foundation of risk regulation and policy making, as well as new information on OSHA and the White House regulatory oversight process.
$18.50

**College Choice in America**
*Charles F. Manski and David A. Wise*
Based on a survey of nearly 23,000 high school seniors, the authors explore the various factors—family income, race, and scholastic aptitude scores—that influence that final decision.
$20.00

**The Control of Resources**
*Partha Dasgupta*
This book examines the issues of resource utilization management and pollution control. Using real-world examples to illustrate difficult concepts, the author pays special attention to critical issues in environmental decision-making.
$18.50

**Competition in the Investment Banking Industry**
*Samuel L. Hayes, III, A. Michael Spence, and David Van Praag Marks*
A major contribution to understanding the investment banking industry, this book provides a comprehensive analysis of the current competitive structure of the industry.
$20.00

**Monetarist Perspectives**
*David Laidler*
Here is a clear and thoughtful introduction to the current literature of monetary economics and macroeconomics, as well as a general view of the macroeconomy in which recession and inflation are to be interpreted.
$20.00

*Now in paperback*

**Equality, The Third World, and Economic Delusion**
*P.T. Bauer*
"Bauer's arguments are brilliant and devastatingly marshalled...[The book] is studded, start to finish, with pure gems of insight and wisdom.—*National Review*
$7.95

**HARVARD UNIVERSITY PRESS**
79 Garden Street,
Cambridge, MA 02138

## *New from The World Bank*

### Telecommunications and Economic Development

*Robert J. Saunders, Jeremy J. Warford, and Bjorn Wellenius*

In this careful, comprehensive examination of the costs and benefits of telecommunications investment, the authors cite examples from both industrialized and developing nations to support their findings in six key areas: benefit estimation and accrual, the impact of investment, regulatory, tariff, and other policies on the benefits of telecommunications expenditures, the benefits versus the costs of telecommunications expansion, the economic analysis of telecommunications investment, and the effects of sector organization and management on sector efficiency. $29.50

*Published for The World Bank by*
**The Johns Hopkins University Press**
Baltimore, Maryland 21218

---

***From the International Monetary Fund . . .***

## Exchange Rate Regimes and Policy Interdependence

| | |
|---|---|
| Flexible Exchange Rates and Interdependence | RUDIGER DORNBUSCH |
| Economic Structure and Policy for External Balance | WILLIAM H. BRANSON |
| Exchange-Rate Regimes and European-U.S. Policy Interdependence | ALEXANDER K. SWOBODA |
| Policy Interdependence from a Latin American Perspective | RICARDO H. ARRIAZU |

The Conference on Exchange Rate Regimes and Policy Interdependence was held in Washington on August 31, 1982 under the joint sponsorship of the International Monetary Fund and the National Bureau of Economic Research. The papers and proceedings have been published in the March 1983 issue of *Staff Papers*, the Fund's quarterly journal. The price of this issue is US$4.00 (special price of US$2.00 for university libraries, faculty, and students). A library bound edition is available for US$8.00.

*Advice on payment in currencies other than the U.S. dollar will be given on request.*

For information and to place orders, please write to:

Publications Unit
International Monetary Fund
Box E-116
Washington, D.C. 20431 U.S.A.

# ECONOMICALLY SOUND AT COLUMBIA AND NYU

## THE LIMITS OF EUROPEAN INTEGRATION

*Paul Taylor.* A study of the prospects for European economic integration which explores the resistance of many nations to the loss of authority in central areas. *The Political Economy of International Change, John Gerard Ruggie, General Editor.* 320 pp., $27.50

## CITY MONEY

### POLITICAL PROCESSES, FISCAL STRAIN, AND RETRENCHMENT

*Terry Clark and Lorna Ferguson.* Based on a comparative study of 62 Standard Metropolitan Statistical Areas, Clark and his colleagues provide the most comprehensive analysis of urban fiscal strain yet available. 384 pp., $32.00 cl, $14.00 pa

## MONETARIST, KEYNESIAN, AND NEW CLASSICAL ECONOMICS

*Jerome L. Stein.* Examines the disagreements between these schools by developing a general macrodynamic model and deriving a set of statistical hypotheses for each school and then critically examining them. 220 pp., *New York University Press* $25.00

## SELECTED ECONOMIC WRITINGS OF ABBA P. LERNER

*David C. Colander, Editor.* Brings together all of Lerner's significant writings on elasticities, the measurement of monopoly power, consumer surplus, the Marshall-Lerner conditions, and many other topics. *New York University Series in Selected Economic Writings.* 752 pp., $65.00

## EAST-WEST TRADE AT THE CROSSROADS

*Robert V. Roosa, Michiya Matsukawa and Armin Gutowski.* An examination of the realistic possibilities for East-West economic relations which calls for an enlightened understanding of the long-term benefits of East-West trade. *A Trilateral Commission Report.* 144 pp., *New York University Press* $15.00

## NATION, STATE, AND ECONOMY

### CONTRIBUTIONS TO THE POLITICS AND HISTORY OF OUR TIMES

*Ludwig von Mises, Leland B. Yeager, Translator.* The first English translation of von Mises' notable 1919 work which identified "statist trends" as the cause of the economic and military ruin of Austria and Germany. *The Institute for Humane Studies Series in Economic Theory.* 248 pp., *New York University Press* $25.00

## AUCTIONS, BIDDING, AND CONTRACTING

### USES AND THEORY

*Richard Engelbrecht-Wiggans, Martin Shubik, and Robert M. Stark, Editors.* Papers written by a broad variety of scholars provide a state-of-the-art overview of the subject both through surveys and in-depth expositions of current topics of interest. *Studies in Game Theory and Mathematical Economics.* 336 pp., *New York University Press* $31.50

## SOVIET ECONOMIC DEVELOPMENT

### SECOND EDITION

*Raymond Hutchings.* Thoroughly revised, with a new section on statistical information concerning recent developments and an entirely new chapter on innovations in Soviet industry. 368 pp, *New York University Press* $32.50 cl, $11.50 pa

## SELECTED ECONOMIC WRITINGS OF W. ARTHUR LEWIS

*Mark Gersovitz, Editor.* The first collection of all of the significant writings of the Nobel laureate on economic development, trade, aid, and other topics. *New York University Series in Selected Economic Writings.* 832 pp. $65.00

***Send for our free Political Science catalog.***

Send check or money order to Dept. JN at the address below, including $2.00 per order for postage and handling.

**COLUMBIA UNIVERSITY PRESS**
136 South Broadway, Irvington, N.Y. 10533

# Books that matter are Basic . . .

## Industrial Renaissance

### Producing a Competitive Future for America

**William J. Abernarthy, Kim B. Clark, and Alan M. Kantrow**

Three leading analysts argue that the current crisis in American industry results less from external factors, like energy costs, than from the failure of American managers to adjust to the realities of today's global markets. In this controversial new book, they explain how the very nature of competition has changed, what demands these changes place on management, and what problems they hold out for the future. "Addresses the key economic issues for the U.S. in the 1980s. . . . Invaluable." —*Steven C. Wheelwright, Stanford University* $19.00

## The Open Economy and the World Economy

### A Textbook in International Economics/John Williamson

Here at last is a textbook in international economics that is truly international, that discusses theory and policy from the perspective of small as well as large economies, and that does not limit itself to specifically U.S. problems. "A very successful text. Skillfully blending theory with facts, Williamson addresses the open economy problems of both industrialized and developing countries while at the same time providing a very accessible account of the relevant theory." —*Rudiger Dornbusch, M.I.T.* $23.95

## The Deindustrialization of America

### Plant Closings, Community Abandonment, and the Dismantling of Basic Industry

**Barry Bluestone and Bennett Harrison**

Business is taking capital out of this country's basic industries at an unprecedented rate. Two leading economists show how and where this "deindustrialization" is taking place and suggest ways to put our economic policies on a more rational footing. "Corporate investment policies have gone astray, exporting our jobs and wrecking our communities. This important book can help us toward a more decent path to economic recovery." —*David M. Gordon, New School for Social Research* $19.95

## Your Money and Your Life

### A Lifetime Approach to Money Management/Robert Z. Aliber

Here is a money management guide that ignores fads and gimmicks to concentrate on developing a lifetime strategy to fit individual financial needs. "Most books on investing should carry the warning: This book may be injurious to your financial health. Robert Aliber's low-keyed book is a welcome exception. It uses economics to explain to you the odds. It can give you an edge in the deadly serious game of personal investing." —*Paul A. Samuelson* $15.95

## . . . new books on economics

### Economic Development

**Theory, Policy, and International Policy**

**Ian M.D. Little /A Twentieth Century Fund Book**

Prominent scholar and policy advisor Ian M.D. Little examines failures, evaluates successes, and makes recommendations for the future in this major study of development economics. "A beautiful book to read, even for those who may disagree. Ian Little's marvelous ability to combine penetrating economic analysis with the wisdom of practical experience makes this book a real treat."—*Amartya Sen, Oxford* $22.95

### Structuralist Macroeconomics

**Applicable Models for the Third World/Lance Taylor**

Here for the first time is a comprehensive and coherent macroeconomics for developing countries. It develops simple, useful, and tested models for different groups of developing countries, models that fit real problems that economists in developing countries must grapple with daily. Lucidly written and brilliantly argued, *Structuralist Macroeconomics* makes an important contribution both to macroeconomic theory and to development economics. $18.95

### The Truth About Supply-Side Economics

**Michael K. Evans**

Michael Evans, one of the nation's leading economic forecasters and himself an early supply-sider, here provides the first objective and documented assessment of the failures — and the achievements — of the supply-side experiment. Although supply-siders misunderstood the psychology of the financial markets and grossly underestimated the political obstacles to fiscal reform, Evans explains, these serious failures should not obscure their real achievement in lowering inflation and setting the country on a new course of long-run economic growth. $17.95

### The International Money Game

**Fourth Edition, Revised and Expanded/Robert Z. Aliber**

When this book first appeared in 1973, it was widely hailed as the best—and most entertaining—introduction to the arcane mysteries of international finance on the market. This revised and expanded fourth edition has been thoroughly updated on topical developments and recent trends. "It's no joke to make a good read out of floating currencies, sliding parities, crawling pegs, gliding rates, Eurodollars and all their unsexy kin, but Professor Aliber, with a nice sense of irony, has done just that."—*The Economist* $15.50 cloth, $8.95 paper

**Basic Books, Inc.** 10 East 53rd Street, New York, N.Y. 10022

# Economist-Ph.D.

Argonne National Laboratory, one of the foremost scientific organizations, seeks a Ph.D. in Economics with several years of outstanding research accomplishments at a University or an organization engaged in economics research. Selected candidate will be expected to assume a leadership role in planning and developing future laboratory programs in resource-related areas. Program areas include biomass resource assessments, strategic materials studies, and energy policy studies.

Responsibilities require at least five years research experience in resource economics or energy technology assessment, with a record of outstanding achievement in one of these areas.

Candidate should have thorough understanding of the principles of welfare economics and resource economics and their application to contemporary energy and natural resource policy issues.

Ideal candidate is one who has demonstrated capability to initiate and conduct original high quality research and has the stature to motivate and lead other professional staff. Well defined written/oral communications skills and the ability to interact effectively with individuals in other disciplines are required.

We offer superior benefits as well as excellent potential for professional recognition. Interested persons should send credentials to Mr. Walter D. McFall, Personnel Division, Box D-EES-3887-94, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439.

An Equal Opportunity/
Affirmative Action Employer

ANNOUNCING THE INTRODUCTION OF THE

## Review of Industrial Organization

Edited by:

Stanley E. Boyle – University of Arkansas at Little Rock
P. David Qualls – University of Tennessee

- The editorial policy of the RIO encourages analytical contributions for both established and beginning researchers.
- The RIO will operate as an open forum for conflicting public policy issues.
- Empirical papers with a solid analytical basis will be welcomed.

All correspondence should be sent to Stanley E. Boyle, Donaghey Professor of Economics, College of Business Administration, University of Arkansas at Little Rock, 33rd and University, Little Rock, AR 72204. Annual Subscription, $25.00 per year, four issues per year. The subscription rate for institutional buyers is $40. The RIO will be published by the University of Arkansas – Little Rock.

Contact Professor Qualls immediately regarding papers related to AEA and SEJ meetings. Manuscript review charge, $15.00 for RIO.

## Reason in Human Affairs

*Herbert A. Simon*. A 1978 Nobel laureate acclaimed for his work on decision-making processes in economic organizations examines the question: What can reason (or more broadly, thinking) do for us and what can't it do? He explores the nature and limits of human reason, comparing and evaluating the major theoretical frameworks that have been erected to explain reasoning processes. He then applies the theory of bounded rationality to social institutions and human behavior, pointing out the problems created by limited attention span—human inability to deal with more than one difficult problem at a time. He concludes that we must recognize the limitations on our capabilities for rational choice, and pursue goals that, in their tentativeness and flexibility, are compatible with those limits. $10.00

*Order from your bookstore, please*

**Stanford University Press**

This issue of the *Review* reports the proceedings of the May 1982 International Research Seminar sponsored by the Chicago Board of Trade.

Five research studies dealing with various options and futures market topics were presented. Each presentation was followed by the critique of an academic and industry member. General discussion of the audience—industry professionals, academic experts, and government officials—completed the sessions.

The *Review* is published at least three times a year by the Chicago Board of Trade in cooperation with the Chicago Board of Trade Foundation.

## Review of Research in Futures Markets

Volume 2 Number 1 1983 Chicago Board of Trade

**Grain Exports, Futures Markets, and Pricing Efficiency**
*Neilson C. Conklin*, Colorado State University
**Information Dissemination Uncertainty, the Continuity of Trading, and the Structure of International Futures Markets**
*Terry A. Marsh*, Massachusetts Institute of Technology, *Robert I. Webb*, Office of Management and Budget and University of Southern California
**Portfolio Strategies Using Treasury Bond Options and Futures**
*Dennis W. Draper*, University of Southern California, *James W. Hoag*, University of California, Berkeley
**Futures Prices and Expected Future Spot Prices**
*Robert W. Kolb*, Emory University, *James V. Jordan*, Georgia Institute of Technology, *Gerald D. Gay*, Georgia State University
**An Analysis of the Impact of Marking-to-Market in Hedging with Treasury Bond Futures**
*Joanne Hill, Thomas Schneeweis, Robert Mayerson*, University of Massachusetts, Amherst

*Subscription price*

$45.00 per volume
$20.00 per issue

*Write or telephone*

Chicago Board of Trade
Literature Services
LaSalle at Jackson
Chicago, IL 60604
312 435 7208/7209

**AMERICAN ECONOMIC ASSOCIATION**

# 1983 Annual Membership Rates

**Membership includes:**

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular member with rank of assistant professor or lower or annual incomes of $15,350 or less . . . . . . $32.00
- Regular member with rank of associate professor or annual incomes of $15,350/$25,600 . . . . . . . . . . . . $38.40
- Regular member with rank of full professor or annual income above $25,600 . . . . . . . . . . . . . . . . . . . . . $44.80
- Junior member (available to registered students for three years only). Student status must be certified by your major professor or school registrar . . . . . . . . . . . . . . . . . . . . . $16.00
- In Countries other than the U.S.A., Add $9.20 to cover postage.
- Family member (second membership without publications; two or more living at same address) . . . . $ 6.40

Please begin my issues with:

☐ **March**  ☐ **June** (Includes *Papers and Proceedings*)  ☐ **September**  ☐ **December**

First Name and Initial ________ Last Name ________ Suffix ________

Address Line 1 ________

Address Line 2 ________

Address Line 3 ________

City ________ State or Country ________ Zip/Postal Code ________

PLEASE TYPE OR PRINT INFORMATION ABOVE; PLEASE SEND CHECK OR MONEY ORDER PAYABLE IN U.S. DOLLARS. CANADIAN AND FOREIGN PAYMENTS MUST BE IN THE FORM OF A U.S. DOLLAR DRAFT ON A NEW YORK BANK.

Endorsed by (AEA member) ________

**Below for Junior Members Only**

I certify that the person named above is enrolled as a student at ________

________

Authorized Signature ________

PLEASE SEND WITH PAYMENT TO:

**AMERICAN ECONOMIC ASSOCIATION**
**1313 21ST AVENUE SOUTH, SUITE 809**
**NASHVILLE, TENNESSEE 37212**
**U.S.A.**

# CSWEP

## *The Committee on the Status of Women in the Economics Profession*

**CSWEP**, an arm of the American Economic Association, publishes a thrice-yearly newsletter to let you know all the news - good or bad - about women in the economics profession.

The newsletter carries items about:

*research on sex role issues*

*new publications*

*hiring (or non-hiring) of women economists*

*conferences for or about women*

*what other professional women's groups are doing to further women's interests in their disciplines*

*developments in government and industry*

**CSWEP** speaks up on behalf of women economists in hiring, research and governmental policies.

**CSWEP** represents women's point of view in the committee work of the American Economic Association. It makes an annual report to AEA on the status of women economists.

**CSWEP** is a presence at annual meetings of the AEA and of the regional economics associations. It sponsors sessions at these meetings, where research by and about women can get an audience.

**CSWEP** publishes a ROSTER OF WOMEN ECONOMISTS for purposes of communication and job placement. Dues-paying members receive the latest roster, which lists women economists by and where they teach or work, by their specialty in economics, and by city, as well as alphabetically.

**CSWEP** is the voice of women economists when coalitions of professional women join to advance sex equality in professional life.

To become a dues-paying member of **CSWEP**
send this with a check for $15 (tax deductible) made out to **CSWEP** to:
**CSWEP**, Box W-M, Economics Department,
University of Maryland, College Park, Md. 20742

NAME ____________________

MAILING ADDRESS ____________________

CITY, STATE, ZIP ____________________

Check here if currently an AEA member ☐

# INDEX OF ECONOMIC ARTICLES

prepared under the auspices of

***The Journal of Economic Literature***

of the

***American Economic Association***

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ Most of the **Index's** volumes also include articles of testimony from selected congressional hearings in government documents published during the year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index.**
- ✓ The 1977 volume contains over 10,500 entries.

**Currently available are:**

| Volume | Year Covered |
|---|---|
| XI | 1969 |
| XII | 1970 |
| XIII | 1971 |
| XIV | 1972 |
| XV | 1973 |
| XVI | 1974 |
| XVII | 1975 |
| XVIII | 1976 |
| XIX | 1977 |
| XX | 1978 (in preparation) |

*an indispensible tool for...*

ECONOMISTS
REFERENCE LIBRARIANS
RESEARCHERS
TEACHERS
STUDENTS
AUTHORS

*Future volumes will be published regularly to keep the series as current as possible.*

**Price: $50.00 per volume (special 30% discount to AEA members)**

Distributed by

*RICHARD D. IRWIN, INC.* Homewood, Illinois
60430

# The Economics Institute

Boulder, Colorado

- PREPARATION FOR MASTERS AND DOCTORAL DEGREE PROGRAMS IN ECONOMICS, BUSINESS, AND ADMINISTRATION.
- POSTGRADUATE DIPLOMA PROGRAMS IN RELATED SPECIALIZATIONS.
- AN ESTABLISHED REPUTATION FOR ACADEMIC EXCELLENCE.

25 years of specialized service in international education.

Sponsored by the American
Economic Association

For further information write:

The Director
Economics Institute
Campus Box 259
University of Colorado
Boulder, Colorado 80309
(303) 492-7337
Telex: 45-0385

Begin making plans to attend the

*Ninety-Sixth*

# Annual Meeting of The American Economic Association

**(in Conjunction with Allied Social Science Associations)**

to be held in

# SAN FRANCISCO, CA.

**Dec. 28-30, 1983**

The Employment Center opens Tuesday, December 27.

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

The 1984 meeting will be held in Dallas, TX, December 28-30.

# International Journal of Industrial Organization

NEW IN 1983

**Managing Editor:**

**Keith Cowling**, *University of Warwick, U.K.*

**Editors:**

**Ken-ichi Imai**, *Hitotsubashi University, Japan*
**Alexis Jacquemin**, *C.R.I.D.E., Catholic University of Louvain, Belgium*
**Dennis C. Mueller**, *International Institute of Management, German Fed. Rep.; University of Maryland, U.S.A.*
**Christian C. von Weizsäcker**, *University of Bern, Switzerland*

**Associate Managing Editor:**

**Paul Stoneman**, *University of Warwick, U.K.*

**Associate Editors:**

**Masahiko Aoki**, *Kyoto University, Japan*
**Raymond De Bondt**, *Catholic University of Leuven, Belgium*
**David Encaoua**, *University of Paris, France*
**Oscar Fanjul**, *University of Madrid, Spain*
**Paul Geroski**, *University of Southampton, U.K.*
**Henry G. Grabowski**, *Duke University, U.S.A.*
**P. Sylos Labini**, *Institute of Economics, Italy*
**Jacques Lesourne**, *Conservatoire National Des Arts Et Métiers, France*
**Frank Mathewson**, *University of Toronto, Canada*
**Tsuruhiko Nambu**, *Gakushuin University, Japan*
**Stephen Nickell**, *London School of Economics and Political Science, U.K.*
**Yoram Peles**, *Hebrew University of Jerusalem, Israel*
**David K. Round**, *University of Adelaide, Australia*
**Steven C. Salop**, *Georgetown University Law Center, U.S.A.*
**Richard Schmalensee**, *Massachusetts University of Technology, U.S.A.*
**Masu Uekusa**, *University of Tokyo, Japan*

**North-Holland Publishing Company**

This new journal is an international venture with strong roots in Europe and Japan, but also with important connexions with the U.S. It will aim at a full coverage of both theoretical and empirical questions within the field of Industrial Organization, broadly defined. As well as covering traditional issues of market structure and performance, the journal will also seek to include articles dealing with the internal organization of firms, all facets of technological change, productivity analysis, and the macro-economic implications of alternative industrial structures. Special attention will be directed at international issues, including industrial structure aspects of trade, investment, technology and development, involving both market and planned economies, and industrialized and industrializing economies.
In collaboration with the various relevant policy-making bodies the journal will also include regular reviews of antitrust and industrial policies in various countries, and within the EEC. A further special feature will be the sponsorship by the journal of occasional symposia on specific issues of topical interest. Some of these will undoubtedly arise from collaboration with the *European Association for Research in Industrial Economics*, out of which many of the ideas for this journal have arisen.

**Contents of the First Issue:**

- MANAGERIALISM REVISITED IN THE LIGHT OF BARGAINING-GAME THEORY, M. Aoki
- CONJECTURES; RATIONALITY AND DUOPOLY THEORY, M. Boyer and M. Moreaux
- HIERARCHIES AND MARKETS: AN EMPIRICAL TEST OF THE MULTI-DIVISIONAL HYPOTHESIS IN WEST GERMANY, J. Cable and M. J. Dirrheimer
- VERTICAL INTEGRATION IN A MONOPOLISTICALLY COMPETITIVE INDUSTRY, A. Dixit
- SOME REFLECTIONS ON THE THEORY AND APPLICATION OF CONCENTRATION INDICES, P. A. Geroski
- EXPERIENCE CURVES AND INDUSTRIAL POLICY, P. E. Hart
- THE PROPENSITY TO PATENT, F. M. Scherer

**Publication Information:**

The INTERNATIONAL JOURNAL OF INDUSTRIAL ORGANIZATION will start publication in 1983. One volume in 4 issues is scheduled for this year.
The institutional subscription price for Volume 1 (1983) is **US $76.00/Dfl. 190.00** including postage and handling. Private subscribers are entitled to a subscription at the reduced rate of **US $34.00/Dfl. 85.00** including postage and handling.

*The Dutch guilder prices are definitive. US $ prices are subject to exchange rate fluctuations.*

**COUPON FOR A FREE COPY**

Send to:
**North-Holland Publishing Company**
**P.O. Box 211, 1000 AE Amsterdam**
**The Netherlands**

For customers in the U.S.A. and Canada:
**Elsevier Science Publ. Co. Inc.**
**Journal Information Center**
**P.O. Box 1663, Grand Central Station**
**New York, NY 10163**

☐ Please send a free specimen copy of the INTERNATIONAL JOURNAL OF INDUSTRIAL ORGANIZATION.

Name ____________________

Professional address ____________________

6124NHa/JNL/ECON

# IUI PUBLISHING

## A NEW CONFERENCE VOLUME

***POLICY MAKING IN A DISORDERLY WORLD ECONOMY***
*edited by G. Eliasson, M. Sharefkin, B-C Ysander*

Did the apparently successful stabilization policy of the 50s and the 60s give us a rigid economic structure that could note cope with the external shocks of the 70s?
Is there a trade-off between successful stabilization and long-run growth?
Sharply increased uncertainty is a fundamental ailment in the world economy. Instead of reducing uncertainty, policy makers have aggraveated the problem through inconsistent ad hoc policy making.
How does an economy adjust to crisis situations and price instabilities? Simular questions are addressed to three different types of econometric models. Quite different answers are given.

INDUSTRIENS UTREDNINGSINSTITUT
Grevgatan 34, S-114 53 STOCKHOLM, Sweden

# OECD publications

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT

## Positive Adjustment Policies: Managing Structural Change

Examines the policy options open to governments for adopting economies to shifts in demand and prices, technological progress, and changing comparative advantage.

| | | | |
|---|---|---|---|
| 03 83 01 1 | February 1983 | 116 p. | $12.50 |

## Job Losses in Major Industries: Manpower Strategy Responses

Describes the scale and consequences of large-scale employment dislocations in such industries as iron and steel, shipbuilding, textiles, and autos in OECD countries. It also analyzes conversion programs designed to reconcile capital mobility with labor protection, and the positive strategies applied in certain countries.

| | | | |
|---|---|---|---|
| 81 83 01 1 | March 1983 | 125 p. | $12.00 |

## External Debt of Developing Countries: 1982 Survey.

Analysis and detailed statistics for the period 1970-1982. Puts the developing country debt problem into perspective.

| | | | |
|---|---|---|---|
| 43 83 01 1 | January 1983 | 122 p. | $12.00 |

## Problems of Agricultural Trade.

Drawing particular attention to the links between trade and national agricultural policy, this study suggests ways of integrating agriculture more fully into the open multilateral trading system with a progressive reduction of protection.

| | | | |
|---|---|---|---|
| 51 82 05 1 | October 1982 | 178 p. | $18.00 |

## Historical Statistics of Foreign Trade, 1965-1980.

Presents total trade of OECD Member countries, volume and value indices, trade by SITC sections, and trade broken down by partner country for the period 1965-1980.

| | | | |
|---|---|---|---|
| 32 82 13 3 | November 1982 | 104 p. | $11.00 |

## World Energy Outlook.

The OECD International Energy Agency's forecasts to the year 2000 for all forms of energy. "... probably the most sophisticated guide in print to the possibilities ahead."—*The Washington Post.*

| | | | |
|---|---|---|---|
| 61 82 08 1 | October 1982 | 474 p. | $45.00 |

*Send orders and requests for information to:*

**OECD PUBLICATIONS AND INFORMATION CENTER**
1750-E Pennsylvania Avenue, N.W.
Washington, D.C. 20006 (202) 724-1857

OECD
OCDE

ECONOMICS

THE SCIENCE OF CHOICE

**LLOYD C. ATKINSON,** *Deputy Assistant Director, Fiscal Analysis Division, Congressional Budget Office; and Adjunct Professor of Economics, University of Maryland*

The principles of Macro and Micro economics are presented in language that is lucid, yet challenging, for the beginning student. The accompanying package includes: Workbook/Study Guide by Dennis Sullivan, Teacher's Manual, Test Bank, Computerized Test Bank, Transparency Masters, and a Semiannual Newsletter.

*More 1982 Editions*

29 AUG 1983

**MACROECONOMICS:**
Analysis and Policy, 4th Edition
**MICROECONOMICS:**
Analysis and Policy, 4th Edition
Lloyd G. Reynolds, Yale University
paperbound/Review Guide and Workbook for each volume

**ECONOMIC ANALYSIS:**
Theory and Application, 4th Edition
S. Charles Maurice and Owen R. Phillips, both of Texas A & M University, and the late C. E. Ferguson

**THE PRACTICE OF COLLECTIVE BARGAINING,**
6th Edition
Edwin F. Beal, University of Oregon, and James P. Begin, Rutgers University

**MONEY, CREDIT, AND ECONOMIC ACTIVITY**
Charles D. Cathcart, Citibank, N.A.

**CASES IN COLLECTIVE BARGAINING AND INDUSTRIAL RELATIONS:**
A Decisional Approach, 4th Edition
Sterling H. Schoen and Raymond L. Hilgert, both of Washington University / paperbound

**MONETARY ECONOMICS AND FINANCIAL MARKETS**
Paul A. Meyer, University of Maryland

**INTERNATIONAL ECONOMICS,** 7th Edition
Peter H. Lindert, University of California-Davis, and Charles P. Kindleberger, Massachusetts Institute of Technology

*When requesting examination copies for adoption consideration, please indicate course title and text presently used.*

**RICHARD D. IRWIN, Inc.** **Homewood, Illinois 60430**

# The American Economic Review

## ARTICLES

O. E. WILLIAMSON
**Credible Commitments: Using Hostages to Support Exchange**

R. H. TOPEL **On Layoffs and Unemployment Insurance**

R. A. HEINER **The Origin of Predictable Behavior**

P. SLOVIC AND S. LICHTENSTEIN
**Preference Reversals**

J. N. BHAGWATI, R. A. BRECHER, AND T. HATTA
**The Generalized Theory of Transfers and Welfare**

J. CARMICHAEL AND P. W. STEBBING
**Fisher's Paradox and the Theory of Interest**

E. P. LAZEAR **A Competitive Theory of Monopoly Unionism**

B. CORNELL **The Money Supply Announcements Puzzle**

C. BULL **Implicit Contracts in the Absence of Enforcement and Risk Aversion**

P. L. MENCHIK AND M. DAVID
**Income Distribution, Lifetime Savings, and Bequests**

B. D. WRIGHT **The Economics of Invention Incentives**

K. G. ABRAHAM **Structural/Frictional vs. Deficient Demand Unemployment**

J. G. MULLIGAN **The Economies of Massed Reserves**

SHORTER PAPERS: R. P. McAfee; J. F. Reinganum; G. M. Grossman and M. L. Katz; J. Bryant; R. Ayanian; J. R. Behrman, R. C. Sickles, and P. Taubman; K. T. Rosen and L. B. Smith; R. Cornes and T. Sandler; W. K. Viscusi; R. G. Chambers; N. E. Bockstael and K. E. McConnell; D. P. Schutte; M. S. Alam; H. Leibenstein; B. J. Caldwell; L. A. Boland; H. Leibenstein; L. De Alessi.

SEPTEMBER 1983

# THE AMERICAN ECONOMIC ASSOCIATION

- Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

- *THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1983, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:
$32.00 for regular members with rank of assistant professor or lower, or with annual income of $15,350, or less;
$38.40 for regular members with rank of associate professor, or with annual income of $15,350 to $25,600;
$44.80 for regular members with rank of full professor, or with annual income above $25,600;
$16.00 for junior members (registered students). Certification must be submitted yearly.
Subscriptions (libraries, institutions, or firms) are $100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.
In countries other than the United States, add $9.20 to cover extra postage.

- Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

- Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

- Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

Founded in 1885

**Officers**

*President*
W. ARTHUR LEWIS
Princeton University

*President-Elect*
CHARLES L. SCHULTZE
The Brookings Institution

*Vice Presidents*
JUANITA M. KREPS
Duke University
EDMUND S. PHELPS
Columbia University

*Secretary*
C. ELTON HINSHAW
Vanderbilt University

*Treasurer*
RENDIGS FELS
Vanderbilt University

*Managing Editor of The American Economic Review*
ROBERT W. CLOWER
University of California-Los Angeles

*Managing Editor of The Journal of Economic Literature*
MOSES ABRAMOVITZ
Stanford University

**Executive Committee**

*Elected Members of the Executive Committee*
ELIZABETH E. BAILEY
Carnegie-Mellon University
ROBERT J. GORDON
Northwestern University
ANN F. FRIEDLAENDER
Massachusetts Institute of Technology
JOSEPH E. STIGLITZ
Princeton University
WILLIAM D. NORDHAUS
Yale University
A. MICHAEL SPENCE
Harvard University

*EX OFFICIO Members*
WILLIAM J. BAUMOL
Princeton University and New York University
GARDNER ACKLEY
The University of Michigan

# GERARD DEBREU

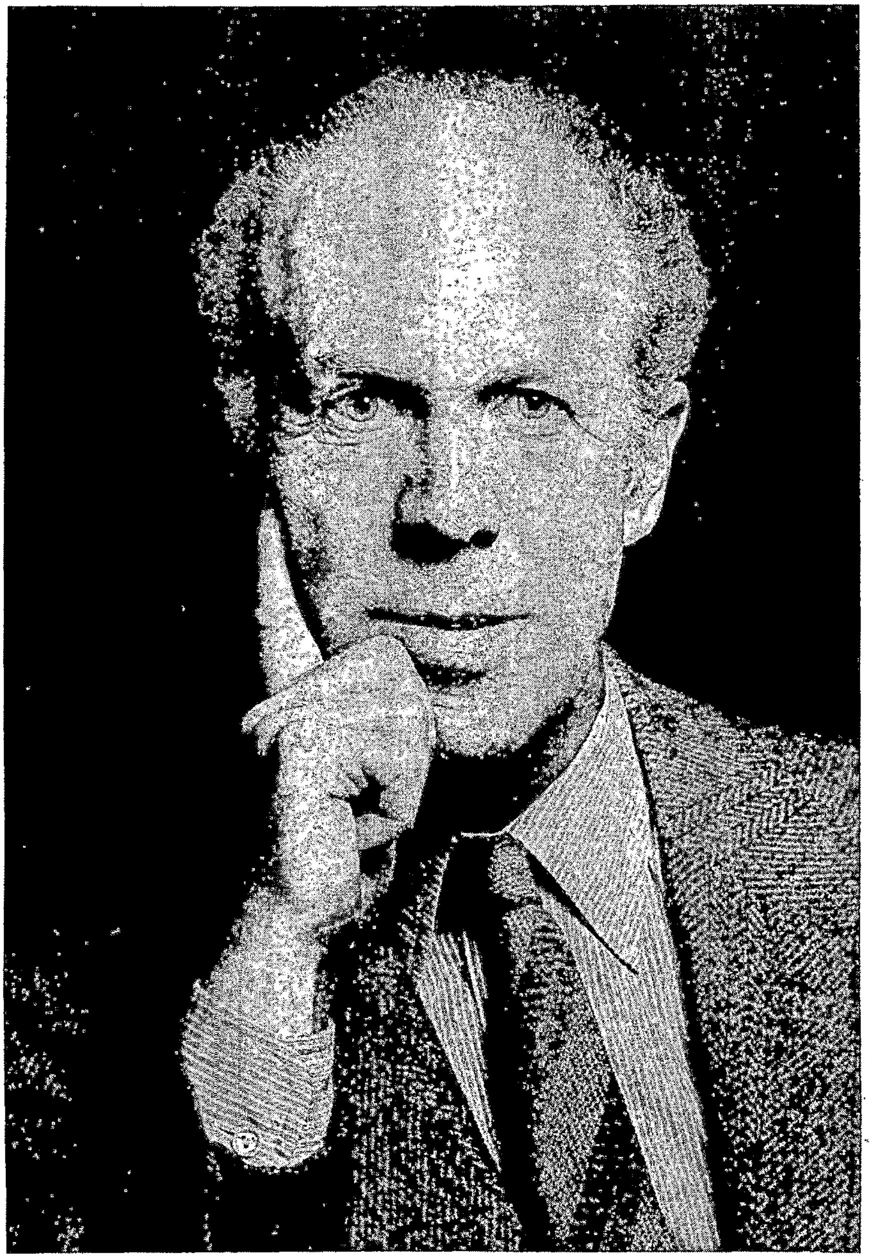
## DISTINGUISHED FELLOW

1982

Gerard Debreu's work has transformed modern mathematical economics. His imprint can be seen everywhere in the field: not only in terms of the important conceptual innovations that he has introduced, but also by his powerful influence on colleagues and students in this country and abroad.

All economists are familiar with his proof of the existence of equilibrium prices in the Walrasian model of a competitive economy. His monograph *Theory of Value* is a classic work in mathematical economics, written with an elegance of exposition and purity of argument which has rarely been approached by other authors in this field. Debreu has made fundamental contributions to the study of the relationship between the core and the competitive equilibrium when the number of participants in the economy tends to infinity; he has played an important role in introducing into economic analysis the techniques of measure theory which have been used to analyze an economy in the presence of an infinite number of agents. Debreu has, as well, clarified the question of uniqueness of the competitive equilibrium, both locally and globally, by means of a judicious adaptation of the methods of differential topology. More recently, he has completed a line of investigation initially begun by Sonnenschein, by demonstrating that market excess demand functions are arbitrary, aside from the most minimal conditions.

In this great variety of intellectual enterprises, Debreu has shown the most remarkable ability to formulate problems at the very heart of a complex issue—problems whose solutions provide an intellectual illumination which extends far beyond the original work. He has been a source of stimulation and the inspiring mentor to a generation of economic theorists.



Gerard Debreu

# THE AMERICAN ECONOMIC REVIEW

ROBERT W. CLOWER
Managing Editor

JOHN G. RILEY
Associate Editor

WILMA ST. JOHN
Production Editor

THERESA DE MARIA
Assistant Editor

LOIS J. BAGLEY
Editorial Assistant

**Board of Editors**
GEORGE A. AKERLOF
G. O. BIERWAG
THOMAS F. COOLEY
PATRICIA DANZON
RONALD G. EHRENBERG
H. E. FRECH III
JACK HIRSHLEIFER
LAURENCE J. KOTLIKOFF
FREDERIC S. MISHKIN
SHERWIN ROSEN
RICHARD SCHMALENSEE
E. ROY WEINTRAUB

•Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

•Copyright © American Economic Association 1983. All rights reserved.

**September 1983**

**VOLUME 83, NUMBER 4**

## Articles

Credible Commitments: Using Hostages to Support Exchange *Oliver E. Williamson* 519

On Layoffs and Unemployment Insurance *Robert H. Topel* 541

The Origin of Predictable Behavior *Ronald A. Heiner* 560

Preference Reversals: A Broader Perspective *Paul Slovic and Sarah Lichtenstein* 596

The Generalized Theory of Transfers and Welfare: Bilateral Transfers in a Multilateral World *Jagdish N. Bhagwati, Richard A. Brecher, and Tatsuo Hatta* 606

Fisher's Paradox and the Theory of Interest *Jeffrey Carmichael and Peter W. Stebbing* 619

A Competitive Theory of Monopoly Unionism *Edward P. Lazear* 631

The Money Supply Announcements Puzzle: Review and Interpretation *Bradford Cornell* 644

Implicit Contracts in the Absence of Enforcement and Risk Aversion *Clive Bull* 658

Income Distribution, Lifetime Savings, and Bequests *Paul L. Menchik and Martin David* 672

The Economics of Invention Incentives: Patents, Prizes, and Research Contracts *Brian D. Wright* 691

Structural/Frictional vs. Deficient Demand Unemployment: Some New Evidence *Katharine G. Abraham* 708

The Economies of Massed Reserves *James G. Mulligan* 725

## Shorter Papers

American Economic Growth and the Voyage of Columbus *R. Preston McAfee* 735
Uncertain Innovation and the Persistence of Monopoly *Jennifer F. Reinganum* 741
Plea Bargaining and Social Welfare *Gene M. Grossman and Michael L. Katz* 749
Government Irrelevance Results: A Simple Exposition *John Bryant* 758
Expectations, Taxes, and Interest: The Search for the Darby Effect *Robert Ayanian* 762
The Impact of Minimum Wages on the Distributions of Earnings for Major Race-Sex Groups: A Dynamic Analysis
*Jere R. Behrman, Robin C. Sickles, and Paul Taubman* 766
The Price-Adjustment Process for Rental Housing and the Natural Vacancy Rate
*Kenneth T. Rosen and Lawrence B. Smith* 779
On Commons and Tragedies *Richard Cornes and Todd Sandler* 787
Frameworks for Analyzing the Effects of Risk and Environmental Regulations on Productivity *W. Kip Viscusi* 793
Scale and Productivity Measurement under Risk *Robert G. Chambers* 802
Welfare Measurement in the Household Production Framework
*Nancy E. Bockstael and Kenneth E. McConnell* 806
Inventories and Sticky Prices: Note *David P. Schutte* 815
Intrafirm Productivity:
  Comment *M. Shahid Alam* 817
  Reply *Harvey Leibenstein* 822
The Neoclassical Maximization Hypothesis:
  Comment *Bruce J. Caldwell* 824
  Reply *Lawrence A. Boland* 828
Property Rights and X-Efficiency:
  Comment *Harvey Leibenstein* 831
  Reply *Louis De Alessi* 843

**Preliminary Announcement of the AEA Program** 846

**Notes** 867

# Credible Commitments: Using Hostages to Support Exchange

*By* OLIVER E. WILLIAMSON*

Credible commitments and credible threats share the following common attribute: both appear mainly in conjunction with irreversible, specialized investments. But whereas credible commitments are undertaken in support of alliances and to promote exchange, credible threats appear in the context of conflict and rivalry.[1] The former involve reciprocal acts designed to safeguard a relationship, while the latter are unilateral efforts to preempt an advantage. Efforts to support exchange generally operate in the service of efficiency; preemptive investments, by contrast, are commonly antisocial. Both are plainly important to politics and economics, but the study of credible commitments is arguably the more fundamental of the two.

Interest in credible threats is much more widespread and the credible threat literature is more fully developed,[2] however, than is the interest and economic literature dealing with credible commitments. This disparity is consistent with the treatment accorded to each in Thomas Schelling's classic essay (1956) on bargaining, where the main emphasis is placed on tactics by which one party can realize an advantage in relation to a rival by credibly "tieing ones hands." But Schelling also, albeit briefly, addresses the matter of promise. He observes in this connection that "Bargaining may have to concern itself with an 'incentive' system as well as the division of gains" (p. 300) and adds in a footnote that the exchange of hostages served incentive purposes in an earlier age (p. 300, fn. 17).

That the study of credible commitments has been relatively neglected is explained by the assumption, common to both law and economics, that the legal system enforces promises in a knowledgeable, sophisticated, and low-cost way. Albeit instructive, this convenient assumption is commonly contradicted by the facts—on which account additional or alternative modes of governance have arisen. Bilateral efforts to create and offer hostages are an interesting and; as it turns out, economically important illustration. Absent a recognition and appreciation for the merits of "private ordering," the suggestion that hostages are used to support contemporary exchange is apt to be dismissed as fanciful. I submit, however, that not only are the economic equivalents of hostages widely used to effect credible commitments, but failure to recognize the economic pur-

---

*Department of Economics, Yale University, New Haven, CT 06520. This paper has benefited from the comments of participants in the Sloan Foundation Workshops in Transaction Cost Economics at the University of Pennsylvania and the University of Minnesota, and from seminar presentations at Oxford University, Williams College, and Yale University. Comments by Dennis Carlton, Frank Easterbrook, Benjamin Klein, Scott Masten, Michael Riordan, David Sappington, Richard Stewart, and Lester Telser, and from the referees of this journal are gratefully acknowledged. Research support from the National Science Foundation is gratefully acknowledged.

[1] It should be noted that I use the terms threat and commitment differently than do Curtis Eaton and Richard Lipsey (1981). They distinguish between empty and credible threats and use the term commitment to refer to the latter. I submit that the language of rivalry is well serviced by reference to threats; and I suggest that the term commitment be reserved to describe exchange. Thus both credible and noncredible threats would be distinguished in assessing rivalry. Similarly, credible and noncredible commitments are distinguished in evaluating exchange. Alliances complicate matters in that these are organized in relation to another party. This could be wholly beneficial, but it need not be. Thus suppliers could form an alliance in relation to buyers, with possible antisocial results. Credible commitments which simultaneously support exchange and promote alliances thus sometimes pose tradeoffs.

[2] Recent applications within economics involve investments in specific capital undertaken for the purpose of impeding new entry (Avinash Dixit, 1979; 1982; Eaton and Lipsey, 1981; Richard Schmalensee, 1981). For a discussion of reputation effects and quasi credibility in the economics literature, see David Kreps and Robert Wilson (1982), Paul Milgrom and John Roberts (1982), and myself (1982b).

poses served by hostages has been responsible for repeated policy error.

The private ordering approach to the study of contract is sketched in Section I. A simple model for assessing the efficiency ramifications of alternative contracts, one of which involves hostages, is developed in Section II. The model demonstrates that the investments made by suppliers are influenced by the incentives experienced by buyers. Incentive complications that reach beyond the model are discussed in Section III. Applications of the argument to unilateral and to bilateral exchange are set out in Sections IV and V. Some evidence bearing on petroleum exchanges and public policy attitudes that relate thereto are examined in Section VI.

## I. Private Ordering

### A. *Contracting Traditions*

Most studies of exchange assume that efficacious rules of law regarding contract disputes are in place and that these are applied by the courts in an informed, sophisticated, and low-cost way. These assumptions are convenient, in that lawyers and economists are relieved of the need to examine the variety of ways by which individual parties to exchange "contract out of or away from" the governance structures of the state by devising private orderings. A division of effort thus arises whereby economists are preoccupied with the economic benefits that accrue to specialization and exchange, while legal specialists focus on the technicalities of contract law.[3]

The "legal centralism" tradition reflects this orientation. It maintains that "disputes require 'access' to a forum external to the original social setting of the dispute...[and that] remedies will be provided as prescribed in some body of authoritative learning and dispensed by experts who operate under the auspices of the state" (Marc Galanter, 1981, p. 1). The facts, however, disclose otherwise. Most disputes, including many of those which, under current rules, could be brought to a court, are resolved by avoidance, self-help, and the like (Galanter, p. 2).

The unreality of the assumptions of legal centralism can be defended by reference to the fruitfulness of the pure exchange model. This is not disputed here. My concern with this tradition is that the law and economics of private ordering have been pushed into the background as a consequence. This is unfortunate, since "In many instances the participants can devise more satisfactory solutions to their disputes than can professionals constrained to apply general rules on the basis of limited knowledge of the dispute" (Galanter, p. 4).[4]

Four distinct, albeit related, literatures within economics[5] have developed over the past decade in which private ordering is expressly or implicitly featured: the incentive compatibility literature (Leonid Hurwicz, 1972); the literature on the economics of internal organization (Ronald Coase, 1937; Kenneth Arrow, 1963, 1974; myself, 1971, 1975, 1979; Benjamin Klein et al., 1978; David Teece, 1982; Eugene Fama and Michael Jensen, 1983); the financial economics literature dealing with bonding (Joseph Stiglitz, 1974; Jensen and William Meckling, 1976; Sanford Grossman and Oliver Hart, 1982); and the study of self-enforcing agreements (Lester Telser, 1981; Klein and Keith Leffler, 1981). The first two of these have been reviewed elsewhere (Hurwicz, 1973; myself, 1981, 1982b). The third takes issue with the Modigliani-Miller theorem that the firm's production plan is independent of its financial structure.[6] The fourth deals with contracting in intermediate product markets and expressly relies on private ordering.

---

[3] Lawyers do not have a monopoly on refining contractual rules. For recent contributions by economists, see Peter Diamond and Eric Maskin (1979) and Steven Shavell (1980). Such an economic approach to contract focuses on the technicalities of legal rules. The approach taken here holds that even refined rules of law are costly to implement, whence private ordering is widely employed.

[4] Galanter elaborates as follows: "The variability of preferences and of situations, compared to the small number of things that can be taken into account by formal rules..., and the loss of meaning in transforming the dispute into professional categories suggest limits on the desirability of conforming outcomes to authoritative rules" (1981, p. 4).

[5] There is also a long legal tradition in which contract as legal rule is disputed. Karl Llewellyn's views regarding "contract as framework" (1931) are especially important. Recent significant contributions include Stewart Macaulay (1963) and Ian Macneil (1974). For a discussion, see my 1979 article.

Telser characterizes a self-enforcing agreement as one which, if "one party violates the terms the only recourse of the other is to terminate the agreement" (p. 27). Contrary to legal centralism, the courts and other third parties are assumed away. Klein and Leffler are explicit on this: "we assume throughout... that contracts are not enforceable by the government or any third party" (p. 616). Commercial contract law in late nineteenth-century Taiwan evidently approximated this condition (Rosser Brockman, 1980). Stewart Macaulay's remarks about the informality of contract in business are likewise in this spirit: "Often businessmen do not feel they have 'a contract'—rather they have 'an order.' They speak of 'cancelling the order' rather than 'breaching our contract'" (1963, p. 61).

To be sure, pure private ordering is extreme. As Robert Mnookin and Lewis Kornhauser state, private ordering invariably operates in "the shadow of the law" (1979).[7] It suffices for my purposes to argue that the incentives of private parties to devise bilateral contractual safeguards is a function of the efficacy of court adjudication, and that this varies with the attributes of transactions. Specifically the courts experience serious disabilities with respect to the transactions of the kinds herein described.

### B. *Some Attributes of this Article*

This article examines self-enforcing agreements in transaction cost terms. Contracting agents are thus assumed to be subject to bounded rationality and, where circumstances permit, are given to opportunism.[8]

Although hostages can have both *ex ante* (screening) and *ex post* (bonding) effects, the *ex post* contract execution consequences are of principal interest here.[9] This is also the focus of the self-enforcing agreement literature. Additionally, like both Telser and Klein-Leffler, the intertemporal contracts of concern here feature both uncertainty and transaction-specific capital. But in other respects there are important differences.

Thus, whereas Telser deals with "a sequence of transactions over time such that the ending date is unknown and uncertain" (p. 30), because any finite sequence of transactions using his model will unravel (p. 29), the transactions that I consider can be (indeed, normally are) finite. Furthermore, the role of transaction specific capital is more explicit and fully developed in this article than in Telser's.

The self-enforcing contracts studied by Klein and Leffler are likewise of indefinite rather than finite duration. The hostage model is further distinguishable from Klein-Leffler in that 1) they deal with quality uncertainty in final goods markets, whereas I assume that quality is known and focus on stochastic demand in intermediate product markets;[10] 2) their "fundamental theoretical result" involves the assurance of quality through the sacrifice of "minimum-cost production techniques" (pp. 618, 628–29), while the hostage model involves no such sacrifice (indeed, the use of hostages to support exchange encourages investment in specific asset technologies which have *lower* expected costs);[11] and 3) suppliers in their model are confronted neither with expropriation nor adaptation hazards, while these are both of concern to me. More generally, the hostage model and imagery have application to a quite different set of contractual circumstances than the analysis of quality assurance contemplates.

---

[6] Grossman and Hart make the interesting distinction between a bonding and a signaling equilibrium: in the former, agents communicate "their endogenous intentions, while the latter involves agents communicating their exogenous characteristics" (p. 110). Put differently, bonding has reference to the incentives of agents at the contract execution stage (an *ex post* condition), whereas signaling involves inferring the directly unobservable *ex ante* attributes of agents, which are fully prespecified. Thus whereas Stephen Ross (1977) uses the debt-equity ratio to signal objective (exogenous) differences in the quality of management, Grossman and Hart use debt to *precommit* managers to a course of action whereby closer *ex post* adherence to profit maximization is induced. Specifically, debt becomes an instrument by which managers place themselves at hazard, in recognition of which the financial market places a higher value on the firm (Grossman and Hart, p. 130). Since managers are assumed to benefit from this higher market valuation (Grossman and Hart, p. 109), they self-consciously accept the hazards of bankruptcy which debt financing poses.

[7] Galanter suggests that a better way to characterize the study of contract is "law in the shadow of indigenous ordering" (p. 23). There is a good deal to be said for this. The main point is that a place for law is properly provided in any comprehensive study of contract.

[8] Whereas previously I have emphasized firm versus market governance, here I focus strictly on market-mediated exchange. The governance issue of interest thus involves choice among alternative contracts.

[9] *Ex ante* screening attributes are briefly examined in an earlier version of this paper (1982a, pp. 6–9). The assessment of a screening equilibrium is complex, however, and is not central to the main argument. See Michael Rothschild and Stiglitz (1976) and John Riley (1979a, b) for a discussion of screening equilibrium issues.

## II. The Hostage Model

The simple hostage model serves to illuminate both unilateral and bilateral exchange, permits the concept of specific capital to be extended beyond earlier uses, and clarifies how costs should be described in assessing exchange. While it is primitive and suggestive, rather than refined and definitive, it serves as a paradigmatic wedge by which the importance of private ordering is exposed and is easily made the vehicle for further analysis.

### A. *Technologies and Costs*

The assessment of alternative contracts will be facilitated by assuming that the product in question can be produced by either of two technologies. One is a general purpose technology; the second is a special purpose technology. The special purpose technology requires greater investment in transaction specific durable assets and, as described below, is more efficient for servicing steady-state demands.

Costs that are highly specific to a transaction have two attributes: they are incurred in advance of the contemplated exchange; and their value in alternative uses, or by alternative users, is greatly reduced.[12] As Klein and Leffler put it, the irreversible, nonsalvageable part of an advance commitment is sunk (p. 619). It is common to think of this as applying to physical plant or accounting costs that are reported as fixed, but this is not the critical distinction. Thus investments in labor (transaction specific human capital) can be highly specific. And many costs that for accounting purposes are reported to be fixed are in fact nonspecific, hence can be recovered (salvaged) by redeployment. Durable but mobile assets such as general purpose trucks or airplanes are illustrations.

The two technologies in question will thus be described in value realization terms. The value that can be realized by redeploying variable and fixed costs will be given by $v$. The nonsalvageable value of advance commitments will be denoted by $k$. The two technologies can thus be described as $T_1$: the general purpose technology, all advance commitments of which are salvageable, the redeployable unit operating costs of which are $v_1$; and $T_2$: the special purpose technol-

---

[10] My discussion of franchising in Section IV.C parallels Klein-Leffler and assumes that quality uncertainty is responsible for a demand externality. The hostage model developed in Section II does not apply directly to this case, but the spirit carries over in three respects: 1) franchisees, like buyers, are given a choice among alternative contracts; 2) the decision to expose specific assets is deliberately taken because this has superior incentive effects; and 3) the sunk cost technology is *more* efficient, which vitiates the inefficiency tradeoff that is central to the Klein-Leffler paper.

[11] Contrary to the argument advanced by Klein and Leffler, total costs need not increase in a quality assurance model because investments are shifted from a reversible (fixed cost) to an irreversible (sunk cost) technology. Thus instead of a general purpose building of nondescript design, the producer could construct a building with a distinctive "signature." The durable investments could be the same, but the alternative value that can be realized from the second building might be much lower. The long-term commitments that are signaled by this second design relieve customers of quality shading hazards, which is the central issue with which Klein and Leffler are concerned.

[12] Klein et al. use the term "appropriable quasi rent" to describe this condition. Use vs. user distinctions are relevant in this connection: "The quasi-rent value of the asset is the excess of its value over its salvage value, that is, its value in its next best *use* to another renter. The potentially appropriable specialized portion of the quasi-rent is the portion, if any, in excess of its value to the second highest-valuing *user*" (p. 298).

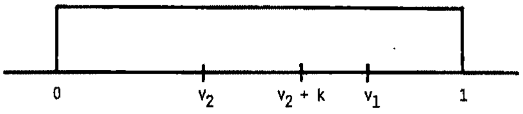


FIGURE 1

ogy, the nonsalvageable value of advance commitments of which are $k$ and the redeployable unit operating costs of which are $v_2$.

### B. *Contracting*

There are two periods. Orders are placed in the first, and production if any occurs in the second. Buyers can either take delivery or refuse it. Demand is stochastic. The gross value to buyers is assumed to be uniformly distributed over the interval [0, 1], and the quantity demanded at every price will be assumed to be constant, which it will be convenient to set equal to unity. Sunk costs, if any, are incurred in the first period. Inasmuch as sunk costs are incurred for certain while the decision to incur redeployable costs is contingent on the buyer's decision to confirm or cancel an order, a choice between technologies is interesting only if $k + v_2 < v_1$. The demand and cost relations are set out in Figure 1.

#### 1. *Net Benefits*

The criterion by which decisions to take or refuse delivery will be evaluated is that of joint profit maximization. Feasibility and/or bureaucratic disabilities aside, vertical integration assuredly accomplishes the joint profit-maximization result. Thus the reference condition for evaluating contracts will be an integrated firm with two divisions, a producing division and a marketing division. The producing division has access to the same two technologies described above, one of which involves specific assets, the other of which does not. Whichever technology is employed, product is transferred between divisions at marginal cost.

That $k + v_2 < v_1$ does not establish that the special purpose technology ($T_2$) is the more efficient. Whether it is or not depends on a net benefit calculation. The expected net benefits of using the general purpose technology ($T_1$) are given by the product of the probability that the integrated firm will decide to produce and the average net benefits that are realized when product is supplied. The integrated firm will decide to produce only if the realized demand price exceeds marginal costs, whence the probability of production under $T_1$ is $1 - v_1$. The mean net benefits during production periods are $(1 - v_1)/2$, whence the expected net benefits for technology $T_1$ are

$$(1) \quad b_1 = (1 - v_1)(1 - v_1)/2 = (1 - v_1)^2/2.$$

The expected net benefits for the specific asset technology ($T_2$) are found similarly. Again, the integrated firm will produce whenever realized demand price exceeds marginal costs. Expected net receipts, however, must be reduced by the amount of the earlier investment in specific assets, $k$, in computing expected net benefits. Thus we have

$$(2) \quad b_2 = (1 - v_2)(1 - v_2)/2 - k$$
$$= (1 - v_2)^2/2 - k,$$

where the first term is the expected excess of revenue over out-of-pocket costs.

The specific asset technology will be selected only if $b_2 > b_1$, which requires that

$$(3) \quad k < (1 - v_2)^2/2 - (1 - v_1)^2/2.$$

#### 2. *Autonomous Contracting*

Assume that the inequality in (3) holds and consider the case of autonomous contracting between a buyer, who services final demand, and a producer, who manufactures the product. Assume that demand and production technologies are as described above. Efficient contracting relations are those that replicate the vertical integration result,[13] namely, 1) select the specific asset technology, and 2) produce and sell product whenever realized demand price exceeds $v_2$. Assume that both parties are risk neutral and that the production side of the industry is competitively organized. Whatever contracting relation is described, producers will be willing to supply if a break-even condition (expressed in expected value terms) can be projected.[14]

---

[13] Here and throughout the remainder of this article I assume that exchange is governed by contract rather than by vertical integration. Vertical integration is thus used merely as a reference condition. That manufacturers do not integrate forward or distributors integrate backward can be explained on a number of

Recall that orders are placed in the first period. Specific assets, if any, are committed in the first period in anticipation of second-period supply. Whether second-period production actually occurs, however, is contingent on demand realizations. Buyers have the option of confirming or cancelling orders in the second period. Consider three contracting alternatives:

I. The buyer purchases specific assets and assigns them to whichever seller submits the lowest bid, $p$;

II. The producer makes the specific asset investment himself and receives a payment of $\bar{p}$ in the second period if the buyer confirms the order but nothing otherwise; and

III. The producer makes the specific asset investment himself and receives $\hat{p}$ from the buyer if the buyer confirms the order, is paid $\alpha h, 0 \leqslant \alpha \leqslant 1$, if the order is cancelled while the buyer pays $\hat{p}$ upon taking delivery and experiences a reduction in wealth of $h$ if second-period delivery is cancelled.

The third scenario can be thought of as one where the buyer posts a hostage that he values in amount $h$, which hostage is delivered to the producer, who values it in amount $\alpha h$ if the order is cancelled.

The producer will break even under contracting relation I if he is compensated in amount $v_2$, which is his out of pocket cost, for each unit demanded. The low bidder will thus offer to supply product for $p = v_2$. Since the buyer's net benefits are maximized if he invests in the specific assets, and since product is transferred on marginal cost terms, this contract replicates the vertical integration relation. Contracts of type I are feasible, however, only if the specialized assets are mobile and the specificity is attributable to physical features (for example, specialized dies). Market procurement can then service the needs of the parties without posing hold-up problems by concentrating the ownership of the specific assets on the buyer (who then assigns them to the low bidder). Inasmuch as the buyer can reclaim the dies and, without cost, solicit new bids should contractual difficulties develop, type I contracts yield an efficient result.[15]

Attention hereafter will be focused on contracts II and III, the assumption being that asset specificity is of the human or dedicated asset kinds (see Part C, below). The autonomous buyer will confirm an order under contract II whenever realized demand price exceeds $\bar{p}$ but not otherwise. The producer will thus break even if $(1-\bar{p})\bar{p} - [(1-\bar{p})v_2 + k] = 0$, whence

$$\bar{p} = v_2 + k/(1-\bar{p}). \tag{4}$$

Product will thus be exchanged at a price that exceeds marginal cost under this contracting scenario.[16] Plainly if $\bar{p} \geqslant v_1$, the

---

grounds—one of which is that there are economies of scope at both stages, but that very different product mixes are needed to realize these scope economies at each stage.

[14] There is no problem in principle in allowing suppliers to extract positive profits as a condition of supply. The salient features of the hostage model are all preserved if, instead of an expected break-even condition, the supplier was assumed to realize expected profits of $\bar{\pi} > 0$ on each contract. Although final demands will be choked off as a consequence, the main features of the contractual argument survive.

[15] This ignores the possibility that suppliers will abuse the dies if ownership resides with the buyer.

[16] Conceivably $\bar{p}$ will exceed $v_1$, in which event the buyer who is contemplating contract II will prefer instead to purchase from sellers who use the general purpose technology. The comparison in the text implicitly assumes that $\bar{p} < v_1$. Also note that a standby technology that can be costlessly switched into and out of the product in question could effectively truncate demand at $v_1$. This would be true if potential middlemen could place orders to take product at $v_1$ from general purpose manufacturers, which orders could be costlessly cancelled (and general purpose assets redeployed) if demands fell below this value. I will arbitrarily assume that this is not feasible. The problem could, however, be reformulated by describing demand as uniformly distributed over the interval 0 to $v_1$, with $v_1$ having measure $1 - v_1$.

buyer is better off to scuttle contract II and purchase instead from producers who utilize the (inferior) variable cost technology $T_1$ (and will break even by supplying product on demand for a price of $v_1$).

The buyer will confirm an order under contract III whenever the realized demand price exceeds $\hat{p}-h$. Let $\hat{p}-h$ be denoted by $m$. The seller will then break even when $(1-m)\hat{p}+m\alpha h-[(1-m)v_2+k]=0$, whence

$$\hat{p}=v_2+(k-m\alpha h)/(1-m). \tag{5}$$

The case where $h=k$ and $\alpha=1$ is one where the buyer gives up wealth in amount of the investment in specific assets in cancellation states and this is delivered to the producer who values it in amount $k$. Under these circumstances, (5) becomes

$$\hat{p}=v_2+k. \tag{5$'$}$$

Since the buyer places an order whenever demand exceeds $m=\hat{p}-h$, this yields the result that $m=v_2$, whence orders will be placed whenever demand exceeds $v_2$, which is the efficient (marginal cost) supply criterion.

The buyer's net benefits under contracting scheme III are

$$b_3=(1-m)\left[\left(m+\frac{1-m}{2}\right)-\hat{p}\right]-mh, \tag{6}$$

where $(1-m)$ is the probability of placing an order, $m+(1-m)/2$ is the expected demand price for all orders that are placed, $\hat{p}$ is the payment in demand confirmation states to the producer, and $h$ is the wealth sacrifice in cancellation states (which occur with probability $m$). Under the assumptions that $h=k$ and $\alpha=1$, this reduces to

$$b_3=(1-v_2)^2/2-k, \tag{6$'$}$$

which is identical to the net benefit calculation for technology $T_2$ under the vertical integration reference condition (see equation (2)).

Accordingly, contracting scheme III accompanied by the stipulations that $h=k$ and $\alpha=1$ replicates the efficient investment and supply conditions of vertical integration. Problems arise, however, if $h<k$ or $\alpha<1$. The disadvantage, moreover, accrues entirely to the buyer—since the seller, by assumption, breaks even whatever contracting relation obtains. Thus although *after* the contract has been made, the buyer would prefer to offer a lesser-valued hostage and cares not whether the hostage is valued by the producer, at the time of the contract he will wish to assure the producer that a hostage of $k$ for which the producer realizes full value ($\alpha=1$) will be transferred in nonexchange states. Failure to make this commitment will result in an increase in the contract price. Thus, whereas producers who are concerned only with *ex ante* screening can tolerate values of $\alpha$ less than one—see the discussion of ugly princesses in Section III.A, below—this is not the case at all when *ex post* opportunism is the concern. If the producer is not indifferent, as between two princesses, each of whom is valued identically by the buyer, the producer's preferences now need to be taken into account.[17]

To summarize, therefore, it can be observed that contract I mimics vertical integration, but only under special asset specificity conditions; contract II is inferior; and contract III yields the vertical integration result if $h=k$ and $\alpha=1$. Furthermore, note that an important feature of contract III is that the buyer takes delivery in all demand states for which realized demand exceeds $m=\hat{p}-h$. Since the supplier is always paid $\hat{p}$ upon execution, the buyer sometimes takes delivery when his realized receipts (upon resale of the product) are less than $\hat{p}$. This does not, however, signal inefficiency, since orders are never confirmed when realized demand price falls below marginal cost ($v_2$). Indeed, it is precisely because of the hostage feature that efficiency is realized and contract III is superior to contract II.

---

[17] Placing an upper bound of unity on $\alpha$ precludes the possibility that the supplier values the hostages more than does the buyer. Potential gains from trade would exist for all hostages for which $\alpha$ exceeds unity. A case for negatively valued hostages could be made in the context of ugly princesses (see Section III, Part A).

### C. *Dedicated Assets*

Three types of asset specificity have been recognized in earlier discussions: site specificity—as where successive stations are located in a cheek-by-jowl relation to each other so as to economize on inventory and transportation expenses;[18] physical asset specificity—such as specialized dies that are required to produce a component; and human asset specificity that arises in a learning-by-doing fashion. A fourth type, previously unremarked but nevertheless important for some transactions, needs to be added: dedicated assets. Such assets represent a discrete investment in plant. Although these assets add to the firm's generalized (as contrasted with special purpose) production capability, the investment would not be undertaken but for the prospect of selling a significant amount of product to a specific customer. As with other types of asset specificity, dedicated assets lose value if employed in alternative uses (or by or to service alternative users). Dedicated assets thus are those that are put in place contingent upon particular supply agreements and, should such contracts be prematurely terminated, would result in significant excess capacity.

## III. Engaging the Supplier

Suppliers are passive instruments in this model. They are indifferent among contracts, since their expected profits are the same (zero) whichever choice the buyer makes. What drives the argument is that buyers can secure better terms only by relieving producers of demand cancellation penalties. Buyers cannot have their cake (product supplied by the efficient technology at a price of $\hat{p}$) and eat it too (cancel without cost).

Inasmuch as optimality is realized if $h=k$ and $\alpha=1$, the ideal hostage would appear to be an offer of generalized purchasing power: money. A security bond in amount $h=k$ would serve this purpose. That the argument does not terminate here is because such an arrangement does not assuredly engage the interests and cooperation of the supplier. Three reasons can be adduced for this condition: contrived cancellation, uncertain valuation, and incomplete contracting. All are a consequence of joining bounded rationality with opportunism.

### A. *Supplier Opportunism*

#### 1. *Contrived Cancellation*

The issue of contrived cancellation has been addressed by Kenneth Clarkson, Roger Miller, and Timothy Muris in their discussion of refusal of the courts to enforce stipulated damage clauses where breach has been deliberately induced (1978, pp. 366–72). Induced breach could arise where a party intentionally withholds relevant information, yet complies with the letter of the contract. Or it might involve perfunctory fulfillment of obligations where more resourceful cooperation is needed (Clarkson et al., pp. 371–72). In either case, induced breach is costly to detect and/or prove (Clarkson et al., p. 371).

This explanation for selective enforcement of liquidated damage clauses has troubled other legal scholars (Richard Posner, 1979, p. 290), but a more satisfactory explanation has yet to be advanced. At the very least, the Clarkson et al. treatment reflects a sensitivity to the subtleties of opportunism—on which account private ordering is more complicated than the bare bones hostage model would suggest. Among other things, the expropriation hazard to which they refer may explain the use of ugly princesses.

Thus suppose that demand uncertainties are negligible, whence order cancellation hazards can be disregarded. Suppose further, however, that buyers differ in credit risk respects, and that producers would, if they could, refuse sales to poor risks. Assuming that the difference between good and poor risks is sufficiently great that a separating equilibrium is feasible,[19] producers could demand hostages (or, put differently, good risks could offer hostages) as a way by which to screen. Given, moreover, that the only use to which hostages are put is as a screen, a value of $\alpha = 0$ would accomplish this purpose without exposing the buyer to an expropriation hazard (based, say, on a legal technicality). Specifically, a king who is known to cherish two daughters equally and is asked, for screening purposes, to post a hostage is better advised to offer the ugly one.

---

[18] Common ownership is the preponderant response to site specificity. Close proximity is desired because of inventory or related processing cost (for example, thermal economy) savings. Once sited, the assets in question are highly immobile—which is to say that the setup and/or relocation costs are great.

[19] See fn. 9.

2. *Uncertain Valuation*

The model assumes that the value of the specific investment ($k$) is well specified. This need not be the case. Indeed, it may be difficult for buyers to ascertain whether the investments made in response to first-period orders are of the amount or of the kind that producers claim. This is not a serious problem if the production side of the market is competitively organized and fly-by-night concerns can be disregarded. Where, however, this cannot be presumed, the possibility that buyers will be expropriated arises. Producers may feign delivery competence (claim to have invested in specific assets in amount $k$ but only committed $k' < k$) and expropriate bonds for which $h = k$ by contriving breach or invoking a technicality.

This hazard is especially great if the producer, who retains possession of the assets for which specificity is claimed, can preserve asset values by integrating forward into the buyer's market upon taking possession of the hostage. Even though the producer is poorly suited to perform successor stage functions, the possession of specialized stage I assets effectively reduces the costs that would otherwise attend de novo stage II entry.

To be sure, the buyer who offers a hostage and recognizes a risk of contrived expropriation will adjust the original terms to reflect this. Specifically, contracts supported by hostages for which expropriation risks are believed to be great will command less than those where these same hazards are believed to be lower. But this is to concede that, absent additional safeguards, neither the transfer of product on marginal cost terms nor the efficient level and kind of investment will assuredly attend contracts of type III. Deeper governance issues than those contemplated by the simple model are evidently posed.

3. *Incomplete Contracts/Haggling*

For the reasons and in the ways described elsewhere (see my 1975 study, pp. 20–36; 91–94), complex contracts are invariably incomplete and many are maladaptive. The reasons are two: many contingencies are unforeseen (and even unforeseeable); and the adaptations to those contingencies that have been recognized and for which adjustments have been agreed to are often mistaken—possibly because the parties acquire deeper knowledge of production and demand during contract execution than they possessed at the outset (Richard Nelson and Sidney Winter, 1982, pp. 96–136). Instrumental gap filling, thus, is an important part of contract execution. Whether this is done easily and effectively, or if instead reaching successive agreements on adaptations and their implementation is costly, makes a huge difference in evaluating the efficacy of contracts.

Thus even if contrived breach hazards could be disregarded, producers who are entirely open and candid about contract execution may nevertheless be in a position to haggle—thereby to expropriate sellers—because contracts are incomplete or maladaptive. Specialized governance structures that have the purpose and effect of promoting harmonious adaptations and preserving the continuity of exchange relations arise in response to this condition. Knowledgeable third parties and reciprocal exposure of specialized assets are two possibilities.

### B. *Protective Governance Structures*

1. *Arbitration*

Institutions that have the capacity to evaluate disputes in a more knowledgeable way than the courts may arise in this way. The parties, for example, may agree to submit disputes over contract execution to arbitrators who have specialized knowledge of the industry. Lon Fuller's remarks concerning procedural differences between arbitration and litigation are instructive:

> ...there are open to the arbitrator... quick methods of education not open to the courts. An arbitrator will frequently interrupt the examination of witnesses with a request that the parties

> educate him to the point where he can understand the testimony being received. The education can proceed informally, with frequent interruptions by the arbitrator, and by informed persons on either side, when a point needs clarification. Sometimes there will be arguments across the table, occasionally even within each of the separate camps. The end result will usually be a clarification that will enable everyone to proceed more intelligently with the case. [1963, pp. 11–12]

Many agreements which, were it not for arbitration, would be regarded as excessively hazardous can, in this way, be reached and implemented.[20]

2. *Reciprocal Exposure*

An alternative way by which to protect contracts against expropriation is to expand the contractual relation. One way of accomplishing this is for buyer and seller to devise a *mutual reliance relation*. Thus suppose that the buyer does not post a hostage as such, but himself invests in specific capital that has value only in conjunction with servicing final demands for the product in question. Assume that these are valued in amount $k''$. The buyer then has the incentive to take delivery as long as realized demand exceeds $\hat{p} - k''$. If $k'' = k$, this yields the marginal cost supply result,[21] and the producer will be satisfied with the buyer's incentives. Or suppose that producer and buyer engage in reciprocal trade. Specifically, suppose that the producer contracts to procure product from the buyer, the supply of which requires the buyer to invest in specific assets in amount $k'''$. Each party to this reciprocal trade will experience appropriate incentives if 1) $k''' = k$, 2) demand variation in the two markets is perfectly correlated, and 3) each party has the option to cancel an order if a cancellation notice is received from his opposite.[22] As discussed in Section V, bilateral trades (reciprocity; swaps) can sometimes be made to approximate these conditions.

## IV. Unilateral Trading

The argument that buyers can affect the terms and manner of supply by offering (or refusing to offer) hostages has ramifications for Robinson-Patman price discrimination and to an understanding of franchising and two-part pricing.

### A. *Robinson-Patman*

The Robinson-Patman Act has been interpreted as an effort "to deprive a large buyer of [discounts] except to the extent that a lower price could be justified by reason of a seller's diminished costs due to *quantity* manufacture, delivery, or sale, or by reason of the seller's good faith effort to meet a competitor's equally low price."[23] Plainly, that $\hat{p}$ is less than $\bar{p}$ in the hostage model has neither quantity nor meeting competition origins. Neither is it contrary to the public interest. Indeed, it would be inefficient and unwarranted for a producer to charge the same price to two customers who order an identical amount of product, but only one of which offers a hostage, if 1) investments in specialized assets are required to support the transactions in question, or 2) if, because of a refusal to make a credible commitment, transactions of the second kind are produced with a general purpose (but high cost) technology.

---

[20] Labor unions can help to assure integrity in contractual relations where workers are asked to accept assignments that involve considerable investments in human capital. Not only can the union intercede on behalf of the worker(s) where an expropriation effort is suspected, but it provides an institutional memory whereby reputation effects can be communicated to successor generations of workers. For both of these reasons, the firm is deterred from attempting expropriation. Setting aside the possibility that unions will attempt to negotiate monopoly wages, perceptive firms will prefer and actively assist in the creation of unions if these serve to attenuate expropriation risks—since otherwise workers may refuse to make (or will need to be bribed to make) mutually beneficial investments in human capital.

[21] The buyer must, of course recover his full costs if he is to place specialized marketing assets at hazard. This will obtain if final demand is uniformly distributed over the interval $k''$ to $1 + k''$ and derived demand is as described earlier.

[22] This last condition protects each against a prisoner's dilemma result.

[23] *FTC v. Morton Salt Co.*, 334 U.S. 37 (1948); emphasis added.

The missing ingredients, plainly, are the differential commitment to buy (as reflected by the willingness to offer hostages) and the differential incentives to breach once hostages have been posted. The confusion is explained by the propensity to employ conventional (steady state) microtheory to the neglect of transaction cost aspects. Rectifying this involves examination of the microanalytics of transactions, with special reference to asset specificity and the hazards thereby posed, and evaluating alternative contracts with respect to a *common reference condition*—prospective break even being a useful standard. Once this is done, a different understanding of many nonstandard or unfamiliar contracting practices, many of which are held to be presumptively unlawful, frequently emerges.[24]

### B. *Franchising*

Klein and Leffler argue that franchisees may be required to make investments in transaction specific capital as a way by which to safeguard the franchise system against quality shading. As Klein puts it, franchisers can better

> ...assure quality by requiring franchisee investments in specific...assets that upon termination imply a capital loss penalty larger than can be obtained by the franchisee if he cheats. For example, the franchiser may require franchisees to rent from them short term (rather than own) the land upon which their outlet is located. This lease arrangement creates a situation where termination can require the franchisee to move and thereby impose a capital loss on him up to the amount of his initial non-salvageable investment. Hence a form of collateral to deter franchisee cheating is created.
> [1980, p. 359]

The arrangement is tantamount to the creation of hostages to restore integrity to an exchange.

This logic notwithstanding, the use of hostages to deter franchisees from exploiting demand externalities is often regarded as an imposed (top down) solution. Franchisees are "powerless"; they accept hostage terms because no others are available. Such power arguments are often based on *ex post* reasoning. That the use of hostages to support exchange can be and often is an efficient systems solution, hence is independent of who originates the proposal, can be seen from the following revised sequence.[25]

Suppose that an entrepreneur develops a distinctive, patentable idea that he sells outright to a variety of independent suppliers, each of which is geographically dispersed and is assigned an exclusive territory. Each supplier expects to sell only to the population located within its territory, but all find to their surprise (and initially to their delight) that sales are also made to a mobile population. Purchases by the mobile population are based not on the reputation of individual franchisees but on customers' perceptions of the reputation of the system. A demand externality arises in this way.

Thus, if sales were made only to the local population, each supplier would fully appropriate the benefits of its promotional and quality enhancement efforts. Population mobility upsets this; since the costs savings that result from local quality debasement accrue to the local operator while the adverse demand effects are diffused throughout the system, suppliers now have an incentive to free ride off of the reputation of the system. Having sold the exclusive territory rights outright, the entrepreneur who originated the program is indifferent to these unanticipated demand developments. It thus remains for the collection of independent franchisees to devise a correction themselves—lest the value of the system deteriorate to their individual and collective disadvantage.

---

[24] Note that the argument applies only to $\hat{p}$ vs. $\bar{p}$ comparisons in trades where specific assets are involved. The efficiency properties of customer price differentials that do have these origins are not reached by the argument in this paper.

[25] That this is a useful way to pose the franchise issue evolved out of discussions that I had with Jeffrey Goldberg. For a more complete development, see Goldberg's dissertation, 1982.

The franchisees, under this revised scenario, thus create an agent to police quality or otherwise devise penalties that deter quality deterioration. One possibility is to return to the entrepreneur and hire him to provide these services. Serving now as the agent of the franchisees, the entrepreneur may undertake a program of quality checks (certain purchasing restraints are introduced, whereby franchisees are required to buy only from qualified suppliers; periodic inspections are performed). The incentive to exploit demand externalities may further be discouraged by requiring each franchisee to post a hostage and by making franchises terminable.[26]

This indirect scenario serves to demonstrate that it is the *system* that benefits from the control of externalities. But this merely confirms that the normal scenario in which the franchiser controls the contractual terms is not an arbitrary exercise of power. Indeed, if franchisees recognize that the demand externality exists from the outset, if the franchiser refuses to make provision for the externality in the original contract, and if it is very costly to reform the franchise system once initial contracts are set, franchisees will bid less for the right to a territory than they otherwise would. It should not therefore be concluded that perceptive franchisers, who recognize the demand externality in advance and make provision for it, are imposing objectionable *ex ante* terms on unwilling franchisees. They are merely taking steps to realize the full value of the franchise. Here as elsewhere, contracts need to be examined in their entirety.

### C. *Two-Part Pricing*

Victor Goldberg and John Erickson (1982) describe an interesting two-part pricing scheme that they observed in the sale of coke. The producer both sold coke to the calciner, and owned and leased the land upon which the plant of the calciner was built. Inasmuch as the coke was sold for "about one-quarter the current market price of equivalent quality coke" (p. 25), Goldberg and Erickson conjecture that "the rental rate was above the fair market rate and that the contract was designed to ensure that [the calciner] would continue to perform" (p. 25). Assuming that marginal costs are much less than average, such an arrangement can be interpreted as one by which the parties are attempting to strike efficient pricing terms that approximate those of the hostage model.

The pricing of utility services, whereby *ex ante* installation fees are paid by subscribers, also have interesting two-part pricing attributes.[27] The risk that sellers will expropriate buyers upon receipt of advance payment can be mitigated by creating a specialized third party, which for convenience may be referred to as a regulatory commission (Goldberg, 1976). Utilization of utility services can then be priced so as to more nearly approximate marginal cost.

More generally, Goldberg and Erickson conjecture that nonlinear pricing schemes are much more widespread than is commonly believed. They further point out that such arrangements are often very subtle and will require detailed knowledge of contracts to investigate (pp. 56–57).

## V. Bilateral Applications

As indicated, the offer of hostages poses a hazard of expropriation. One way to deter this is to expand the contracting relationship from one of unilateral to bilateral exchange. Credible commitments are signaled without exposing assets to expropriation hazards. Reciprocal trades, especially those that involve product exchanges (swaps), sometimes come about in this way.

### A. *Reciprocity, General*

Reciprocity transforms a unilateral supply relation—whereby $A$ sells $X$ to $B$—into a bilateral one, whereby $A$ agrees to buy $Y$

---

[26] Termination is a credible threat only if the franchisee who cheats on the system bears a capital loss. This is the basic Klein and Leffler message. It would not do, therefore, if the terminated franchisee were permitted to sell the franchise to a highest bidder unless the investment in specific capital took the form of the franchisee's specialized knowledge of the system, and the terminated franchisee were thereafter prohibited from participating in owner, adviser, or employee status.

[27] This possibility was suggested to me by Alvin Klevorick.

from *B* as a condition for making the sale of *X* and both parties understand that the transaction will be continued only if reciprocity is observed. Although reciprocal selling is widely held to be anticompetitive (George Stocking and Willard Mueller, 1957; Harlan Blake, 1973), others regard it more favorably. George Stigler offers the following affirmative rationale for reciprocity:

> The case for reciprocity arises when prices cannot be freely varied to meet supply and demand conditions. Suppose that a firm is dealing with a colluding industry which is fixing prices. A firm in this collusive industry would be willing to sell at less than the cartel price if it can escape detection. Its price can be reduced in effect by buying from the customer-seller at an inflated price. Here reciprocity restores flexibility of prices. [1969, p. 39]

Inasmuch as, however, many industries do not satisfy the prerequisites for oligopolistic price collusion (Posner; myself, 1975, ch. 12) and as reciprocity is sometimes observed among these, reciprocity presumably has other origins as well. Tie breaking is one of these. A second is that reciprocity can have advantageous governance structure benefits. These two can be distinguished by the type of product being sold.

The tie-breaker explanation applies where firm *B*, which is buying specialized product from *A*, asks that *A* buy standardized product from *B* on the condition that *B* meets market terms. Other things being equal, procurement agents at *A* are apt to accede. F. M. Scherer notes that "Most of the 163 corporation executives responding to a 1963 survey stated that their firms' purchases were awarded on the basis of reciprocity only when the price, quality, and delivery conditions were equal" (1980, p. 344).

The more interesting case is where reciprocity involves the sale of specialized product to *B* conditioned on the procurement of specialized product from *B*. The argument here is that reciprocity can serve to equalize the exposure of the parties, thereby reducing the incentive of the buyer to defect from the exchange—leaving the supplier to redeploy specialized assets at greatly reduced alternative value. Absent a hostage (or other assurance that the buyer will not defect), the sale by *A* of specialized product to *B* may never materialize. The buyer's commitment to the exchange is more assuredly signaled by his willingness to accept reciprocal exposure of specialized assets. Defection hazards are thereby mitigated.

Lest the argument be uncritically considered to be a defense for reciprocal trading quite generally, note that it applies only where specialized assets are placed at hazard by *both* parties. Where only one or neither invests in specialized assets, the practice of reciprocity plainly has other origins.[28]

### B. *Exchanges*

Although reciprocal trading among nonrivals may occasionally be justified, the exchange of product among nominal rivals is surely more puzzling and troublesome. Firms that are presumed to be in head-to-head competition ought to be selling product against one another rather than to one another. What explains the reverse?

Several distinctions are useful in considering exchanges. First, trade among rivals—short term or long term, unilateral or bilateral—is feasible only if product is fungible. This is not true for many differentiated goods and services, whence the issue of trade among rivals never arises for these. Second, short-term supply agreements are usefully distinguished from long term. The former may be explained as an "occasional exception," whereby one rival will sell product to another on a short-term, gap-filling basis so as to provide temporary relief against unanticipated product shortfalls (occasioned by either demand or supply changes). Recognizing that the shoe may be on the other foot next time, otherwise rivalrous firms may assist one another for stop-gap purposes. Public policy can presumably recognize merit in such trades and, so long as they lack a pattern, hence do not give rise to a "web of interdependence," will regard these as unobjectionable. Long-term trading among rivals is, however, much less consistent with the notion of effective head-to-head rivalry. At the very least, such arrangements warrant scrutiny.

---

[28] Possible trading objections are discussed by Scherer (pp. 344–45). Another objection is that reciprocity becomes a bureaucratic habit that salesmen and purchasing agents find convenient and that outsiders are thereby disadvantaged in attempting to secure sales. See my 1975 study, pp. 163–64.

Whether there are efficiency incentives for rivals to supply product to one another on a long-term basis turns initially on prospective realization of production cost savings. The realization of production cost savings through long-term trade between rivals requires that economies of scale be large in relation to the size of geographic markets and, if they are, that firm-specific reputation effects extend across geographic market boundaries. The former is obvious since, absent economies of scale, every firm would presumably supply everywhere to its own long-term needs. Where, however, scale economies are significant, each market will support only a limited number of plants of minimum efficient size.

But fungibility and scale economies do not establish that gains from trade will be realized from such sales. This will obtain only if the value of (identical) product sold by rivals exceeds that sold by the local supplier. The issue here is whether valued reputation effects will go unrealized if rivals are unable to secure local product on favorable terms. Firms that possess valued reputations that extend beyond their local market to include distant markets are thus the ones for which long-term supply by rivals will be attractive.[29]

Even supposing that fungibility, scale economy, and reputation effect conditions are satisfied, this merely establishes that *unilateral* long-term trade among rivals can yield economies. A justification for *bilateral* (exchange) agreements is not reached by these arguments. Indeed, the usual defense for exchanges—that inefficient cross-hauling will occur if every firm is required to supply everywhere to its own needs—conveniently suppresses the obvious alternative, which is not zero trade, but rather unilateral long-term trade. Failure to address these matters directly and demonstrate wherein exchanges enjoy comparative institutional advantages over more standard and familiar forms of unilateral trade presumably explains the suspect or hostile attitude with which exchanges are typically regarded. The argument that emerges from this paper is that bilateral exchanges offer prospective advantages over unilateral trade *if* the resulting exposure of transaction specific assets effects a credible commitment without simultaneously posing expropriation hazards.

The type of specific asset that is placed at hazard by unilateral long-term trade, but which a reciprocal long-term exchange agreement serves to protect, is that of a dedicated asset. Recall that dedicated assets were described as discrete additions to generalized capacity that would not be put in place but for the prospect of selling a large amount of product to a particular customer. Premature termination of the contract by the buyer would leave the supplier with a large excess of capacity that could be disposed of only at distress prices. Requiring buyers to post a bond would mitigate this hazard, but only by posing another: the supplier may contrive to expropriate the bond. More generally, the interests of the supplier in adapting efficiently to new circumstances are not fully engaged. Reciprocal trading supported by separate but concurrent investments in specific assets provides a mutual safeguard against this second class of hazards. The hostages that are thereby created have the interesting property, moreover, that they are *never exchanged*. Instead, each party retains possession of its dedicated assets should the contract be prematurely terminated.

The usual argument that exchanges are justified because they avoid costly cross-hauling does not get to these issues and, by itself, is not an adequate justification for widespread use of exchanges. Were it only that transportation cost savings were realized, unilateral trading would suffice. Indeed, petroleum firms should be expected to create a central exchange in which supplies and demands were brought into correspondence by an auctioneer. Firms would end up selling to each other only by accident in these circumstances. Where dedicated assets are exposed, however, *the identity of the parties clearly matters*. Trades of this kind will not go through an auction market but will be carefully negotiated between the parties. Reciprocity in these circumstances is thus a device by which the continuity of a specific trading relation is promoted with risk attenuation effects.

---

[29] Reputation effect valuations may be illusory or real. Those that are real take the form of customer convenience (billing, contracting) or assured knowledge of product characteristics.

## VI. Petroleum Exchanges

"The task of linking concepts with observations demands a great deal of detailed knowledge of the realities of economic life" (Tjalling Koopmans, 1957, p. 145). The phenomenon of petroleum exchanges has puzzled economists for a long time. It routinely comes up in antitrust cases and investigations. The 1973 case brought by the United States Federal Trade Commission against the major petroleum firms maintained the view that exchanges were instrumental in maintaining a web of interdependencies among these firms, thereby helping to effect an oligopolistic outcome in an industry that was relatively unconcentrated on normal market structure criteria.[30] The more recent study on *The State of Competition in the Canadian Petroleum Industry* likewise regards exchanges as objectionable.[31] The Canadian Study, moreover, produces documents—contracts, internal company memoranda, letters, and the like—as well as deposition testimony to support its views that exchanges are devices for extending and perfecting monopoly among the leading petroleum firms. Such evidence on the details and purposes of contracting is usually confidential and hence unavailable. But detailed knowledge is clearly germane—and sometimes essential—to a microanalytic assessment of the transaction cost features of contract.

### A. *The Evidence from the Canadian Study*

Volume V of the Canadian Study deals with the refining sector. Arguments are advanced and supporting evidence is developed that interfirm supply arrangements permit the major refiners to perfect oligopolistic restrictions in the following four respects:[32] 1) valuable knowledge about investment and marketing plans of rivals are disclosed by such agreements (p. 56); 2) leading firms are able to control lesser firms by exercising discretionary power through the terms of exchange (pp. 49–50); 3) competition is impaired by conditioning supply on the payment of an "entry fee" (pp. 53–54); and 4) exchange agreements impose limits on growth and supplementary supply (pp. 51–52).

The first two of these fail to pass scrutiny of the most rudimentary comparative institutional kind. Thus assuming that trade between rivals is efficient and that unilateral supply agreements (if not exchange) will be permitted, the objectionable information disclosures attributed to exchanges would presumably continue—since investment and marketing plans will be unavoidably disclosed in the process. Accordingly, evaluated in comparative institutional terms, the information disclosure objection is properly regarded as an objection to long term trade of any kind. Exchanges are not uniquely culpable.

The suggestion that exchanges are anticompetitive because they permit firms to realize bargaining advantages is similarly misplaced. The correct view is that firms should always be expected to realize such bargaining advantages as their positions lawfully permit. Absent a showing that exchanges are different from unilateral trades in bargaining respects, this objection is properly disregarded also.

---

[30] *FTC v. Exxon et al.*, Docket No. 8934 (1963).

[31] Robert J. Bertrand, Q. C., Director of Investigation and Research, Combines Investigation Act, coordinated the eight-volume study, *The State of Competition in the Canadian Petroleum Industry* (Quebec, 1981). All references in this paper are to Vol. V, *The Refining Sector*. This study will hereinafter be referred to as the Canadian Study.

[32] The Canadian Study contends that "a close examination of the interest of the [major refiners] and their actions shows that refining arrangements were meant to restrict competition. The collection of information, the intent to control lesser firms, the imposition of an 'entry fee,' the use of restrictions on downstream growth are not characteristics that would be expected normally from a competitive market" (Vol. V, p. 76).

The entry fee and marketing restraint objections are more substantial, however, and warrant elaboration.

1. *Entry Fees*

The entry fee objection to exchanges is that this has foreclosure consequences. That such fees are required as a precondition for trade, or at least the sale of product at favorable prices, is set out in the Canadian Study as follows:

> Evidence of an understanding that a fee relating to investment was required for acceptance into the industry can be found in the following quotation from Gulf: "We do believe that the oil industry generally, although grudgingly, will allow a participant who has paid his ante, to play the game; the ante in this game being the capital for refining, distributing and selling products" (Document #71248, undated, Gulf). The significance of the quotation lies equally in the notion that an "entry fee" was required and in the notion that the industry set the rules of the "game." The meaning of the "entry fee" as well as the rules of the "game" as understood by the industry can be found in the actual dealings between companies where the explicit mention of an "entry fee" arises. These cases demonstrate the rules that were being applied—the rules to which Gulf was referring. Companies which had not paid an "entry fee", that is, companies which had not made a sufficient investment in refining capacity or in marketing distribution facilities would *either not be supplied or would be penalized in the terms of the supply agreement.*
> [pp. 53–54, emphasis added]

2. *Marketing Restraints*

The Canadian Study notes that exchanges were made conditional on growth and territorial restraints and regards both as objectionable. The Imperial-Shell exchange agreement, under which Imperial supplied product to Shell in the Maritimes and received product in Montreal, is cited in both connections.

> The agreement between Imperial and Shell, originally signed in 1963, was renegotiated in 1967. In July 1972, Imperial did this because Shell had been growing too rapidly in the Maritimes. In 1971/72, Imperial had expressed its dissatisfaction with the agreement because of Shell's marketing policies. Shell noted:
>
> "There [sic] [Imperial's] present attitude is that we have built a market with their facilities, we are aggressive and threatening them all the time, and they are not going to help and in fact get as tough as possible with us" (Document #23633, updated, Shell). [Vol. 5, p. 51]

Imperial renewed the agreement with Shell only after imposing a price penalty if expansion were to exceed "normal growth rates" and furthermore stipulated that "Shell would not generally be allowed to obtain product from third party sources" to service the Maritimes (p. 52).

Gulf Oil likewise took the position that rivals receiving product under exchange agreements should be restrained to normal growth: "Processing agreements (and exchange agreements) should be entered into only after considering the overall economics of the Corporation and should be geared to providing competitors with volumes required for the normal growth only."[33] It furthermore sought and secured assurances that product supplied by Gulf would be used only by the recipient and would not be diverted to other regions or made available to other parties (p. 59).

### B. *Interpretations*

These practices are subject to several interpretations. One is that the entry fees and marketing restraints are both anticompetitive. A second is that efficiency purposes are arguably served, especially by the former. A third is that there are mixed effects.

---

[33] The Canadian Study (p. 59) identifies the source as Document #73814, January 1972, Gulf.

### 1. *The Inhospitality Tradition*

The two polar contracting traditions for evaluating nonstandard or unfamiliar contracting practices are the common law tradition, and the antitrust or inhospitality tradition. Donald Turner makes reference to both of them in expressing his views about vertical market restrictions: "I approach territorial and customer restrictions not hospitably in the common law tradition, but inhospitably in the tradition of antitrust law."[34] Thus whereas contractual irregularities are presumed to serve affirmative economic purposes under the common law tradition, a deep suspicion of anticompetitive purposes is maintained by the antitrust (or inhospitality) tradition.

The inhospitality tradition is supported by the widespread view that economic organization is technologically determined. Economies of scale and technological nonseparabilities explain the organization of economic activity within firms. All other activity is appropriately organized by market exchanges. Legitimate market transactions will be mediated entirely by price; restrictive contractual relations signal anticompetitive intent.

The authors of the Canadian Study are evidently persuaded of the merits of this tradition. Long-term trade among rivals of any kind is suspect. And exchanges, which represent an irregular if not unnatural contracting form, are especially objectionable. Not only do exchanges facilitate information disclosure and permit bargaining strength, but they are used punitively against nonintegrated independents who, because they have not paid an entry fee, are denied product on parity terms. Furthermore, the marketing restraints that are associated with exchanges are patently offensive.

### 2. *An Efficiency Assessment*

Unlike the inhospitality tradition, the transaction cost approach is in the common law tradition. A comparative institutional orientation (Coase, 1964) is maintained. "Defects" are thus objectionable only where superior feasible alternatives can be described. Inasmuch as the information disclosure and bargaining concerns raised by the authors of the Canadian Study continue under unilateral trading, these are set aside and attention is focused on other matters.

(a) Entry Fees. The entry fee issue is a matter of special interest to this paper. Long-term exchange agreements permit firms to secure product in geographic markets where own-production is not feasible because economies of scale are large in relation to their own needs. The amount of product in question may nevertheless be substantial. Firms with whom exchange agreements are reached will thus construct and maintain larger plants than they otherwise would. Specific investments in dedicated assets are made as a consequence of such agreements.

Were it that supply agreements were of a unilateral kind and the buyer was unable or unwilling to offer a hostage, contracts of type II would presumably be negotiated—whence the trading price would be $\bar{p} = v_2 + k/(1-\bar{p})$. If, instead, the contract is extended to include bilateral rather than unilateral trade, the contract is converted to one of type III. Although exchange agreements stipulate the physical flows of product, the effective price is $\hat{p} = v_2 + k$, which is less than $\bar{p}$. Moreover, the parties have the incentive to exchange product so long as realized demand price in both regions exceeds $v_2$,[35] which is the marginal cost supply criterion. Assuming that demands in the two regions are highly correlated, the parties will normally reach common decisions on the desirability of trade.[36]

---

[34] The quotation is attributed to Turner by Stanley Robinson, 1968, N.Y. State Bar Association, Antitrust Symposium, p. 29.

[35] This assumes common costs, which condition will normally be approximated in exchanges of product between firms within a single country where factor prices are very similar.

[36] The possibility that the contract will drift out of alignment nevertheless needs to be recognized. Should one of the firms in an exchange agreement operate much closer to its capacity limits than the other, the latter party would incur much higher costs of termination than would the former. Recognition of this may explain

(b) Marketing Restraints. The supply and growth restraints discussed by the Canadian Study can be looked at in three ways. First, these can be viewed as a means by which to protect the exchange agreement against unilateral defection. Second, such restraints may serve strategic market division purposes. Third, restraints may serve to regularize markets. These are not mutually exclusive.

Only the first of these purposes is consonant with an efficiency interpretation. The argument here is that marketing restraints help to preserve symmetrical incentives. Such symmetry could be upset if one of the firms were to receive product in its deficit region from third parties. Such a firm might then be in a position to play one supplier off against the other. Or symmetry could be placed under strain if one party were to receive product from the other such that it began to grow "in excess of normal"—in which event it might be prepared to construct its own plant and scuttle the exchange agreement. Marketing restraints which help to forestall such outcomes encourage parties to participate in exchanges that might otherwise be unacceptable.

### 3. *A Mixed View*

Monopoly explanations are commonly advanced when economists, lawyers, or other interested observers come across contractual practices that they do not understand. Inasmuch as "we are very ignorant [in this field], the number of ununderstandable practices tends to be very large, and the reliance on a monopoly explanation frequent" (Coase, 1972, p. 67). A rebuttable presumption that nonstandard contracting practices are serving affirmative economic purposes, rather than monopoly purpose, would arguably serve antitrust law and economics better than the inhospitality presumption which, until recently, has prevailed.[37]

The presumption that exchanges have efficiency purposes could be challenged on any or all of three grounds. First, it might be argued that exchanges are merely a clever device by which to deny product to nonintegrated rivals. Refusals to sell to nonintegrated firms on $\bar{p}$ terms would support this contention. (It is plainly unrealistic, however, for buyers that have not made credible commitments to expect to receive product at $\hat{p}$.) Second, the market in question could be shown to have troublesome structural properties. The issue here is whether the requisite preconditions for market power—mainly high concentration coupled with high barriers to entry[38]—are satisfied. A third would be that the preconditions for efficiency are not satisfied. Factors favorable to the efficiency interpretation are the following: the exchange should be of a long-term kind; the amount of product exchanged should represent a significant fraction of plant capacity; and economies of plant scale should be large in relation to the amount of product traded. Exchanges for a small quantity of product where economies of scale are insubstantial are much more problematic.

To be sure, exchanges might simultaneously service efficiency and anticompeti-

---

why "during the renegotiation of a reciprocal purchase/sale agreement covering Montreal and the Maritimes," Shell noted that Imperial advised them that "they were not satisfied with the extent of Shell's investment in the Maritimes" (p. 54). In addition to the investment in refining in Montreal, which Shell interpreted as an investment "by exchange" in the Maritimes, Imperial wanted Shell to make direct investment in a Maritime distribution network (p. 54). Shell observed in this connection that although it had made no significant investment of its own in the Maritimes, "we have invested in Montreal and by exchange invested in the Maritimes so we have paid an entrance fee, although we have not paid for distribution network." The Canadian Study (p. 54) identifies the source as Document #23633, updated, Shell.

[37] To be sure, this is an oversimplification. Antitrust, has been loath to declare contractual constraints to be per se illegal. It came perilously close to taking this step in *U.S. v. Arnold Schwinn & Co.*, 388 U.S. 365 (1967), however. The prevailing enforcement view toward contractual restraints in the 1960's is accurately characterized as inhospitable.

[38] There is growing agreement that the structural preconditions that must be satisfied before claims of strategic anticompetitive behavior are seriously entertained are very high concentration coupled with barriers to entry (my 1977 article, pp. 292–93; P. L. Joskow and A. K. Klevorick, 1979, pp. 225–31; Janusz Ordover and Robert Willig, 1981, pp. 307–08).

tive purposes. Here as elsewhere, where tradeoffs are posed, they need to be evaluated.

## VII. Concluding Remarks

The study of contract in both law and economics has mainly emphasized legal rules and technicalities. Such an orientation is supported by the implicit assumption that the courts "work well." Whether they work well or poorly, however, requires a comparative institutional assessment. There is growing awareness that the (comparative) limitations of the courts are more severe than the legal centralism tradition admits.

The severity of these limitations is not uniform but varies with the circumstances. A discriminating approach to the study of contract will necessarily make provision for this. Specifically, if different transactions have different governance needs, these will be expressly recognized. Accordingly, the study of contract is appropriately extended from legal rules to include an assessment of alternative governance structures, of which the courts are only one. Of special interest in this connection is the use of bilateral governance structures (private ordering) to implement nonstandard contracts where the adaptation and continuity needs of the parties are especially great.

This paper is an effort to deepen the understanding of private ordering. The central points are these:

1) Hostages: Contrary to the prevailing view that hostages are a quaint concept with little or no practical importance to contemporary contracting, the use of hostages to support exchange is widespread and economically important. But hostage creation is only part of the story. Expropriation hazards and prospective maladaptation conditions also need to be considered. Complex governance structures, of which reciprocal trading is one, arise in response to such conditions.

2) Asset Specificity: The organization of economic activity is massively influenced by the degree to which the transactions under examination are supported by assets that are specific to the parties. This paper 1) reaffirms the basic proposition that governance structures need to be matched to the underlying attributes of transactions in a discriminating way if the efficiency purposes of economic organization are to be realized, 2) extends the scope of asset specificity to include dedicated assets, and 3) establishes that, as between two buyers, one of whom posts a hostage in support of specific asset investments by suppliers while the other does not, suppliers will offer better terms to the former, *ceteris paribus*.

3) Microanalytics: The relevant unit of analysis for studying exchange relations of the kinds discussed in this article is the transaction. Assessing transactions and assigning them to governance structures in a discriminating (mainly transaction cost economizing) way requires much more microanalytic knowledge of economic activity and organization than is customary within economics. Empirical work will necessarily reflect this.[39] Price and quantity of course remain relevant, but the contractual devices by which prices are made to track costs, the manner in which adaptations are effected, and the safeguards that are provided are not only germane but are sometimes decisive.

4) Contracting in its Entirety: Not every transaction poses defection hazards, and it may not be possible to safeguard all that do. Where the potential hazards that beset contracts are evident to the parties from the outset, however, studies of contracts and of contracting institutions arguably start "at the beginning." This has ramifications for assessing the importance of the prisoners' dilemma and for understanding the administration of justice.

(a) Prisoners' Dilemma: The benefits of cooperation notwithstanding, the achievement of cooperation is widely thought to be frustrated by the relentless logic of the prisoners' dilemma. To be sure, it has always been evident that defection can be deterred if payoffs are appropriately altered. But this stratagem is held to be infeasible or is otherwise dismissed—on which account the dilemma persists or appeal is made to "exogenous norms of cooperative behavior [that are] adhered to by the actors" (Albert Hirschman, 1982, p. 1470). I submit that the feasibility of crafting superior *ex ante* incentive structures warrants more attention. A leading reason for its neglect is because the study of the institutions of contract has occupied such a low place on the research agenda. Subtle incentive features that are incorporated in nonstandard contracting practices have gone undetected as a consequence of this nonchalance—on which account the practical significance of the prisoners' dilemma to the study of exchange has been vastly exaggerated.

---

[39] Examples of microanalytic studies of contract include myself (1976), Thomas Palay (1981), Goldberg and Erickson (1982), and Scott Masten (1982).

(b) Justice: The notion that hostages are demanded as a condition for supplying product on favorable terms has the appearance of an arbitrary exercise of power: the stronger party "demands" a hostage from the weaker, who accedes it because it has no other choice. In fact, a comparative institutional assessment of contractual alternatives discloses that efficiency purposes are often served by hostages and that it is in the mutual interest of the parties to achieve this result. Not only can producers be induced to invest in the most efficient technology, but buyers can be induced to take delivery whenever demand realizations exceed marginal cost. More generally, contracts need to be examined *in their entirety*, with special attention to their governance features. Principles of justice or competition that look at the relation between the parties at the execution stage without examining the *ex ante* bargaining relation are at best incomplete and are frequently mistaken.[40] Parties to a contract should not expect to have their cake (low price) and eat it too (no hostage).

## REFERENCES

**Arrow, Kenneth J.**, "Control in Large Organization," *Management Science*, September 1963, *10*, 397–408.

______, *Limits of Organization*, New York: W. W. Norton, 1974.

**Blake, Harlan M.**, "Conglomerate Mergers and the Antitrust Laws," *Columbia Law Review*, March 1973, *73*, 555–92.

**Brockman, Rosser H.**, "Commercial Contract Law in Late Nineteenth Century Taiwan," in Jerome Alan Cohen et al., eds., *Essays on China's Legal Tradition*, Princeton: Princeton University Press, 1980, 76–136.

**Clarkson, Kenneth W., Miller, Roger L. and Muris, Timothy J.**, "Liquidated Damages v. Penalties," *Wisconsin Law Review*, 1978, 351–90.

**Coase, Ronald H.**, "The Nature of the Firm," *Economica*, November 1937, *4*, 386–405; reprinted in George J. Stigler and Kenneth E. Boulding, eds., *Readings in Price Theory*, Homewood: Richard D. Irwin, 1952.

______, "The Regulated Industries: Discussion," *American Economic Review Proceedings*, May 1964, *54*, 194–97.

______, "Industrial Organization: A Proposal for Research," in Victor R. Fuchs, ed., *Policy Issues and Research Opportunities in Industrial Organization*, Vol. 3, New York: National Bureau of Economic Research, 1972, 59–73.

**Diamond, Peter A. and Maskin, Eric**, "An Equilibrium Analysis of Search and Breach of Contract," *Bell Journal of Economics*, Spring 1979, *10*, 282–316.

**Dixit, Avinash**, "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, *10*, 20–32.

---

[40] Robert Nozick's views on justice are apposite: "whether a distribution is just depends upon how it came about. In contrast, *current-time-slice* principles of justice hold that the justice of a distribution is determined by how things are distributed (who has what)" (1975, p. 153). What he refers to as the current-time-slice approach to justice neglects *ex ante* bargaining and evaluates justice in terms of outcomes alone. Upon realization that justice is administered in this way, initial bargains will be struck on different terms than they would if the parties were given assurance that the complete contract would be subject to review in evaluating the merits of a contracting relation. Two difficult issues nevertheless remain if the comprehensive bargain orientation to justice is adopted: the initial distribution of resources; and the competence of the parties to evaluate complex contracts. The relative importance of these varies with the circumstances.

______, "Recent Developments in Oligopoly Theory," *American Economic Review Proceedings*, May 1982, *72*, 12–17.

**Eaton, B. Curtis and Lipsey, Richard G.**, "Capital, Commitment, and Entry Equilibrium," *Bell Journal of Economics*, Autumn 1981, *12*, 593–604.

**Fama, Eugene F. and Jensen, Michael C.**, "Separation of Ownership and Control," *Journal of Law and Economics*, June 1983, *26*.

**Fuller, Lon L.**, "Collective Bargaining and the Arbitrator," *Wisconsin Law Review*, January 1963, 3–46.

**Galanter, Marc**, "Justice in Many Rooms: Courts, Private Ordering, and Indigenous Law," *Journal of Legal Pluralism*, No. 19, 1981, 1–47.

**Goldberg, Jeffrey**, "A Theoretical and Econometric Analysis of Franchising," draft, doctoral dissertation, University of Pennsylvania, 1982.

**Goldberg, Victor P.**, "Regulation and Administered Contracts," *Bell Journal of Economics*, Autumn 1976, *7*, 426–48.

______ **and Erickson, John E.**, "Long-Term Contracts for Petroleum Coke," Department of Economics Working Paper Series No. 206, University of California-Davis, September 1982.

**Grossman, Sanford J. and Hart, Oliver D.**, "Corporate Financial Structure and Managerial Incentives," in John J. McCall, ed., *The Economics of Information*, Chicago: University of Chicago Press, 1982, 107–40.

**Hirschman, Albert**, "Rival Interpretations of Market Society: Civilizing, Destructive, or Feeble?," *Journal of Economic Literature*, December 1982, *20*, 1463–84.

**Hurwicz, Leonid**, "On Informationally Decentralized Systems," in C. B. McGuire and R. Radner, eds., *Decision and Organization*, Amsterdam: North-Holland 1972, 297–336.

______, "The Design of Mechanisms for Resource Allocation," *American Economic Review Proceedings*, May 1973, *63*, 1–30.

**Jensen, Michael and Meckling, William**, "Theory of the Firm: Managerial Behavior, Agency Costs, and Capital Structure," *Journal of Financial Economics*, October 1976, *3*, 305–60.

**Joskow, P. L. and Klevorick, A. K.**, "A Framework for Analyzing Predatory Pricing Policy," *Yale Law Journal*, December 1979, *89*, 213–70.

**Klein, Benjamin**, "Transaction Cost Determinants of 'Unfair' Contractual Relations," *American Economic Review Proceedings*, May 1980, *70*, 356–62.

______, **Crawford, R. A. and Alchian, A. A.**, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

______ **and Leffler, K. B.**, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, *89*, 615–41.

**Koopmans, Tjalling C.**, *Three Essays on the State of Economic Science*, New York: McGraw-Hill, 1957.

**Kreps, David M. and Wilson, Robert**, "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, *27*, 253–79.

**Llewellyn, Karl N.**, "What Price Contract?—An Essay in Perspective," *Yale Law Journal*, May 1931, *40*, 704–51.

**Macaulay, Stewart**, "Non-Contractual Relations in Business," *American Sociological Review*, February 1963, *28*, 55–70.

**Macneil, Ian R.**, "The Many Futures of Contract," *University of California Law Review*, May 1974, *67*, 691–816.

**Masten, Scott**, "Transaction Costs, Institutional Choice, and the Theory of the Firm," unpublished doctoral dissertation, University of Pennsylvania, 1982.

**Milgrom, Paul and John Roberts**, "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, August 1982, *27*, 280–312.

**Mnookin, Robert H. and Kornhauser, Lewis**, "Bargaining in the Shadow of the Law: The Case of Divorce," *Yale Law Journal*, March 1979, *88*, 950–97.

**Nelson, Richard R. and Winter, Sidney G.**, *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

**Nozick, Robert**, *Anarchy, State, and Utopia*, New York: Basic Books, 1975.

**Ordover, Janusz A. and Willig, Robert D.,** "An Economic Definition of Predatory Product Innovation," in S. Salop, ed., *Strategic Views of Predation*, Washington: Federal Trade Commission, 1981, 301–96.

**Palay, Thomas,** "The Governance of Rail-Freight Contracts: A Comparative Institutional Approach," unpublished doctoral dissertation, University of Pennsylvania, 1981.

**Posner, Richard A.,** "Some Uses and Abuses of Economics in Law," *University of Chicago Law Review*, Winter 1979, *46*, 281–306.

**Riley, John G.,** (1979a) "Informational Equilibrium," *Econometrica*, March 1979, *47*, 331–53.

______, (1979b) "Noncooperative Equilibrium and Market Signaling," *American Economic Review Proceedings*, May 1979, *69*, 303–07.

**Ross, Stephen,** "The Determination of Financial Structure: The Incentive Signaling Approach," *Bell Journal of Economics*, Spring 1977, *8*, 23–40.

**Rothschild, Michael and Stiglitz, Joseph,** "Equilibrium in Competitive Insurance Markets," *Quarterly Journal of Economics*, November 1976, *80*, 30–49.

**Schelling, Thomas C.,** "An Essay on Bargaining," *American Economic Review*, June 1956, *46*, 281–306.

**Scherer, F. M.,** *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

**Schmalensee, Richard,** "Economies of Scale and Barriers to Entry," *Journal of Political Economy*, December 1981, *89*, 1228–38.

**Shavell, Steven,** "Damage Measures for Breach of Contract," *Bell Journal of Economics*, Autumn 1980, *11*, 446–90.

**Stigler, George J.,** in *President's Task Force on Productivity and Competition*, reprinted in CCH *Trade Regulation Reporter*, No. 419, June 24, 1969.

**Stiglitz, Joseph,** "Incentives and Risk Sharing in Sharecropping," *Review of Economic Studies*, June 1974, *41*, 219–57.

**Stocking, George W. and Mueller, Willard F.,** "Business Reciprocity and the Size of Firms," *Journal of Business*, April 1957, *30*, 73–95.

**Teece, David J.,** "Economics of Scope and the Scope of the Enterprise," *Journal of Economic Behavior and Organization*, September 1980, *1*, 223–45.

______, "A Transaction Cost Theory of the Multinational Enterprise," unpublished manuscript, 1982.

**Telser, L. G.,** "A Theory of Self-Enforcing Agreements," *Journal of Business*, February 1981, *53*, 27–44.

**Wachter, Michael L. and Williamson, Oliver E.,** "Obligational Markets and the Mechanics of Inflation," *Bell Journal of Economics*, Autumn 1978, *9*, 549–71.

**Williamson, Oliver E.,** "The Vertical Integration of Production: Market Failure Considerations," *American Economic Review Proceedings*, May 1971, *61*, 112–23.

______, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

______, "Franchise Bidding for Natural Monopolies—in General and with Respect to CATV," *Bell Journal of Economics*, Spring 1976, *7*, 73–104.

______, "Predatory Pricing: A Strategic and Welfare Analysis," *Yale Law Journal*, December 1977, *87*, 284–340.

______, "Transaction-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, *22*, 233–61.

______, "The Modern Corporation: Origins, Evolution, Attributes," *Journal of Economic Literature*, December 1981, *19*, 1537–68.

______, (1982a) "Mitigating Contractual Hazards: Using Hostages to Support Exchange," Discussion Paper No. 126, Center for the Study of Organizational Innovation, University of Pennsylvania, April 1982.

______, (1982b) "Antitrust Enforcement: Where It has Been; Where It is Going," in John Craven, ed., *Industrial Organization, Antitrust, and Public Policy*, Boston: Kluwer-Nijhoff Publishing, 1982, 41–68.

# On Layoffs and Unemployment Insurance

By ROBERT H. TOPEL*

The problem studied in this paper is how current systems of providing and financing unemployment insurance (*UI*) affect the private decisions that generate unemployment. It is widely recognized that the provision of *UI* affects the search strategies of jobless individuals by raising reservation wages, thus increasing the average duration of unemployment spells.[1] In various forms, almost all empirical work on the impact of *UI* has studied this incentive. More recently, the focus of theoretical research has shifted to the role of *UI* in affecting the joint decisions of workers and firms that generate transitions *to* unemployment. Here the emphasis is on current methods of financing benefits—via partially "experience rated" payroll taxes on individual employers—and the incentives that these methods provide toward increasing the incidence of temporary layoff unemployment.[2] In terms of what is known about magnitudes of effects, this role of experience rating is probably the most important unresolved empirical issue in *UI* research. It is the main concern in this paper.

That current methods of financing unemployment insurance subsidize unemployment is undeniable; as it turns out, experience rating in determining employer taxes is normally incomplete, so that in almost all cases, the value of benefits received by unemployed workers exceeds their incremental cost to firms. The empirical magnitude of this wedge, which often equals or even exceeds the total money value of benefits paid out, can provide a powerful incentive toward increased layoffs. Despite this fact, however, previous research has been severely limited by the absence of any reliable measure of the experience rating subsidy to unemployment that is both relevant for individual workers, and that can be used in empirical analysis.[3]

My point of departure in this paper is in measuring the extent of subsidization that is implied by the structures of *UI* financing systems in the United States. Differences in these structures across states provide the empirical leverage needed to identify the incentive effects of the *UI* subsidy. The estimates reported below, based on a large sample of individuals from the 1975 Annual Demographic File of the *Current Population Survey* (*CPS*), indicate that incomplete experience rating may account for as much as 30 percent of all spells of temporary layoff unemployment. Additionally, most of the impact of *UI* on layoffs is accounted for by this subsidy; nonsubsidized benefits are found to have an insignificant impact on layoff decisions.

Appealing to the theory of job search, the duration of unemployment spells has played a prominent role in previous empirical research on the effects of *UI*. The incentives

---

*Department of economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637. I acknowledge support from the Center for the Study of the Economy and the State at the University of Chicago, and the U.S. Department of Labor. I have benefited from the comments of David Coppock, Joe Hotz, Sherwin Rosen, George Stigler, and Finis Welch, and of seminar participants at Chicago, Rochester, SUNY-Stony Brook, University of California-Los Angeles, and Yale University. Joe Tracy provided excellent research assistance.

[1] The empirical literature on duration effects of *UI* is large and continues to grow. For useful surveys of empirical estimates, see Finis Welch (1977) or Daniel Hamermesh (1979). Relatively recent contributions are Nicholas Kiefer and George Neumann (1979) and Steven Nickell (1979).

[2] The main reference is Martin Feldstein (1976). See also the discussions in Martin Baily (1976), and my article with Welch (1980). Temporary layoffs may account for as much as 50 percent of unemployment spells among workers who have separated from their previous jobs (quits, discharges, and layoffs). This share is strongly countercyclical, and is larger in manufacturing than elsewhere. See the discussion in Feldstein (1975), and my (1982a) article.

[3] For example, Feldstein (1978) measures *UI* incentives by the ratio of available benefits to disposable weekly earnings. Nothing can be inferred from this variable about the role of experience rating and *UI* subsidization in affecting unemployment. Only Frank Brechling (1981) has attempted to tie experience rating attributes of states to measures of labor turnover.

that operate in the present context are quite different, however, and theory indicates that *UI* will affect *both* the probabilities that individuals enter unemployment (the frequency of spells, via layoff decisions) and that they leave unemployment (the duration of spells, via search and rehire decisions). The empirical methodology developed below estimates the effects of *UI*, and other variables, on each of these transitions. Consistent with theory, the estimates indicate that the *UI* subsidy increases unemployment by changing both the frequency and the expected duration of layoff spells. It is important to note that this methodology adjusts for the fact that *UI* may induce a greater number of short spells, which could actually reduce the mean duration of unemployment in any sample.

The paper is organized as follows. The first section develops the empirical foundation for the problem analyzed here, focusing on the structure of *UI* systems and methods of financing in the United States, and on the incentives implied for the joint employment decisions of workers and firms. Section II describes the data and the empirical strategy used to estimate *UI* effects on unemployment, and Section III reports the econometric evidence.

## I. The Institutional Setting

Among unemployed individuals who have separated from their previous jobs (i.e., excluding entrants to the labor force), temporary layoffs and discharges consistently dominate quits, accounting for nearly 90 percent of this category in an average year.[4] Since quits are the major reason for *UI* ineligibility under state laws, it is clear that the large majority of passages to unemployment are, in principle, compensable by the *UI* system. In fact, tabulations of a special questionnaire administered with the May 1976 *CPS* indicate that 75 percent of all temporary layoffs and 70 percent of discharges were either receiving *UI* benefits or had an application to receive them pending. In short, most "involuntary" transitions into unemployment involve the receipt of *UI* benefits for at least some portion of time spent unemployed.

Despite this important connection, not all individuals who may experience unemployment are equally "insured" by the *UI* system. Subject to several qualifying restrictions, a typical state program may pay weekly benefits to an unemployed worker that are equal to half of foregone, before-tax weekly earnings, up to a maximum for a stipulated period. Variations on this basic algorithm are numerous, however, and the proportion of spendable earnings that are replaced by *UI* benefits—the so-called "replacement ratio"—can differ across individuals both within and between state programs. To illustrate the range of these differences, Table 1 shows calculated values of the replacement ratio for hypothetical individuals in several state programs in 1975. Thus, for example, a married worker in Michigan with two dependents and a weekly wage of $100 would have experienced only a 4 percent decline in disposable income had he become unemployed,[5] while a layoff would have reduced the same worker's income by more than 40 percent in California or Florida. Similar differences can be generated within state programs: a tripling of the weekly wage to $300 in Texas implies a 34 percentage point decline in the replacement ratio, while the same experiment in Wisconsin—with a more liberal benefit ceiling and progressive state income taxes—would reduce the replacement ratio only slightly.

With experience-rated *UI* financing, the effects of these differences on the incidence of unemployment, and especially on layoff decisions, are theoretically ambiguous.[6] Con-

---

[4] These estimates are based on tabulations of the *CPS* data for the twelve years, 1968–79. Temporary layoffs are defined here to include individuals who have a definite recall date within thirty days of the survey, plus those whose layoff is of indefinite duration. Discharges are unemployed workers who do not anticipate rehire by their former firms.

[5] Though not shown in the table, a worker in Michigan with three or more dependents would experience a significant *increase* in disposable income during an unemployment spell.

[6] These points are developed formally in the expanded version of this paper, which examines the demand for insurance by workers and the feedback effects of workers' search strategies on the layoff and rehire decisions of firms. See my (1982a) article.

TABLE 1—BENEFIT REPLACEMENT RATIOS FOR QUALIFIED WORKERS IN SELECTED STATE UNEMPLOYMENT INSURANCE PROGRAMS, 1975

| | Weekly Wage | | | |
|---|---|---|---|---|
| | Single | | Married, 2 Dependents | |
| State | $100 | $300 | $100 | $300 |
| California | .63 | .50 | .59 | .45 |
| Florida | .62 | .38 | .59 | .35 |
| Massachusetts | .70 | .45 | .81 | .46 |
| Michigan | .76 | .45 | .96 | .50 |
| New York | .65 | .46 | .69 | .42 |
| Ohio | .62 | .40 | .73 | .41 |
| Texas | .64 | .29 | .61 | .27 |
| Wisconsin | .67 | .63 | .63 | .57 |

*Note*: Replacement ratio is calculated as potential benefit amount under state qualifying provisions as a proportion of after-tax weekly earnings. Marginal tax rates are based on standard deductions under federal and state income tax laws for the indicated family structure. Benefit amounts reflect dependency allowances where applicable.

sider, for example, the case where wages and benefits do not substitute perfectly in worker preferences (workers are risk averse), so that there is a private demand for *UI* as insurance for workers. In the absence of a *UI* system, optimal employment agreements between workers and firms would provide for some form of private unemployment compensation that mitigates workers' income risks. In this context, the incentive effects of publicly administered *UI* depend on 1) who pays for the benefits that workers receive, and 2) the level of required benefit payments relative to what would have existed privately. In general, if *UI* is a binding constraint on employment agreements, so firms must pay benefits above the privately optimal level, then higher benefits will increase the cost of laying off workers and the incidence of unemployment will decline.[7] In contrast, if firms are not liable for any increase in benefits, then *UI* subsidizes unemployment and the incidence of layoffs will increase. Our current systems of financing benefit payments result in a diverse mixture of these offsetting extremes, with firms usually being only partially liable for the benefits their workers receive. Thus, the size and even the sign of *UI*'s effect on unemployment depend on the degree of subsidization that is relevant in any particular case.

The size of this subsidy can be quite important. In the United States, as in no other country, *UI* benefits are financed by taxes on employer payrolls that are related to their individual histories of generating unemployment. As already noted, this system of experience rating is highly imperfect. Yet some employers are more heavily subsidized by imperfect rating than others. To focus on these differences, and to highlight the main empirical issues, Table 2 reports some characteristics of the distribution of insured unemployment and tax liabilities for six major state systems in 1967 and 1978.[8]

An important feature of state programs is that assigned employer tax rates are bounded, so that firms with relatively high average unemployment, who pay the maximum rate, may consistently accumulate deficits of tax liabilities relative to benefit payments. For these employers and their workers, the marginal cost of benefits is zero since an increase in insured unemployment can cause no incremental taxes. The empirical importance of this fact is illustrated in row A of the table, which shows for each state the estimated proportion of total employment that occurred in firms whose accumulated past tax contributions were smaller than benefit withdrawals. Especially in 1967, this proportion is fairly small (averaging about 10 percent of covered employment), yet row B shows that roughly *half* of all *UI* benefits were received by employees of these firms. Evidently, firms that entered these years with negative balances due to high past unemployment had higher than normal unemployment during 1967 and 1978 as well. In fact, rows C and D show that estimated unemployment rates for deficit employers in 1967

---

[7] The same argument implies that firms would have a greater incentive to rehire a worker on layoff in order to avoid the costs of continued benefits. Thus, the duration of spells would be shorter as well. Of course, for workers who are discharged from their previous jobs, so there is no possibility of recall, only search incentives are relevant and average duration of spells would increase.

[8] The data used in these calculations were collected and compiled by Joseph Becker (1972, 1981). Complete data for other states and years are not available.

TABLE 2—SUMMARY STATISTICS FOR SELECTED STATE PROGRAMS: THE DISTRIBUTION OF INSURED UNEMPLOYMENT FOR POSITIVE AND NEGATIVE BALANCE FIRMS, 1967 AND 1978

| | California | | Massachusetts | | Michigan | | New York | | Ohio | | Wisconsin | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1967 | 1978 | 1967 | 1978 | 1967 | 1978 | 1967 | 1978 | 1967 | 1978 | 1967 | 1978 |
| A. Proportion of Total Taxable Wages Paid by Negative Balance Firms[a] | 14.2 | 14.3 | 11.8 | 13.1 | 3.6 | 28.0 | 13.8 | 28.4 | 4.4 | 29.9 | 8.5 | 14.4 |
| B. Proportion of Total Charged Benefits Charged to Negative Balance Firms[a] | 51.8 | 52.5 | 55.3 | 46.0 | 34.8 | 68.1 | 61.6 | 72.7 | 34.2 | 58.1 | 60.8 | 26.9 |
| C. State Insured Unemployment Rate[b] | 3.9 | 3.3 | 2.9 | 3.1 | 2.6 | 3.6 | 2.9 | 3.8 | 1.6 | 2.1 | 2.1 | 2.5 |
| D. Estimated Insured Unemployment Rate for Negative Balance Firms[c] | 14.2 | 12.1 | 13.6 | 10.9 | 25.1 | 8.7 | 12.9 | 9.6 | 12.4 | 4.0 | 15.0 | 4.7 |

[a] *Source*: Becker (1972, p. 112; 1981, p. 83).

[b] Average weekly insured unemployment. *Source*: U.S. Department of Labor, *Handbook of Unemployment Insurance Financial Data* (1978). Data for 1978 are unpublished, and were obtained directly from the Labor Department.

[c] (Row B ÷ Row A) × Row C.

averaged about five times the implied means for positive balance employers, but this difference sharply narrowed as greater numbers of firms achieved deficits by 1978.[9] Thus, the data indicate extensive cross subsidization within self-financing state systems. And since most firms with negative balances pay the maximum allowable tax rates, it follows that the marginal cost of benefits that is relevant for a major portion of insured unemployment is also heavily subsidized.

While these facts document an important degree of cross subsidization in the current *UI* system, nothing is implied by them about the empirical role of *UI* financing in affecting unemployment, since workers in high unemployment sectors are, by construction, subject to larger subsidies. For example, the insured unemployment rate for workers in the construction industry is typically about 10 percent, which is sufficiently high to guarantee that employers in this industry pay the maximum rate and face a full marginal subsidy on benefits. It is clearly not legitimate to infer from this fact that large subsidies cause above-average unemployment in this industry. This type of reverse causality presents a serious inference problem in isolating independent effects of the subsidy, and methodology for dealing with it is developed below. In addition, while these data indicate that a significant portion of unemployment is subject to zero experience rating, they say nothing about the costs and incentives implied for the majority of firms who pay tax rates between the state minima and maxima, and who are experience rated to some degree.[10] This *degree* of rating, which is implied by the structure of state systems, plays an important role in the empirical analysis. Here, I outline the essential features of measuring the degree of experience rating, and relegate technical details to the Appendix. More detailed descriptions of particular experience rating methods appear in my article with Welch.

---

[9] *Ceteris paribus*, this shift implies that a larger proportion of covered employment is subject to zero experience rating. For details and causes of these trends, see my article with Welch (1980).

[10] For the states listed in Table 2, about 75 percent of employment fell in this category. For further evidence, see Becker (1981, Table 24, pp. 164–67).

### A. *The Degree of Experience Rating in Current Financing Systems*

Under federal legislation, states may design their own schemes of benefit financing and payroll taxation so long as assigned tax rates can be justified on the basis of an employer's unemployment history.[11] As with Social Security, these taxes are paid on a taxable wage base per employee that may be chosen by the states.[12] Federal requirements on the form of state systems are quite loose, and as a result there is considerable heterogeneity both in the type of system chosen and, for a given system, in the particular parameters that affect incentives. In this paper, I focus on two general methods of experience rating accounting, the reserve-ratio and benefit-ratio methods, because they are the most common and because their incentives can be parsimoniously summarized for empirical research. Together, these methods apply to more than 80 percent of insured employment in the United States.

I define the degree of experience rating $e$ as the value of marginal tax liabilities paid by employers per dollar of benefits received by workers. Thus, an employer faces incomplete experience rating ($e<1$) if an additional layoff spell generates benefits that have a greater present value than the associated increase in taxes. Empirically, there are two common features of financing systems that assure incomplete rating for most employers. First, as noted above, assigned tax rates are bounded above and below so that taxes are completely insensitive to changes in layoff behavior in some sectors. In these cases, $e=0$. Second, even for rated firms whose tax rates fall between the minimum and maximum in a state, where taxes are sensitive to layoffs, incremental tax liabilities are spread through time. Interest is not charged on these liabilities, so the degree of rating depends on the time profile of tax changes. Thus, even in cases where the full nominal value of benefits is eventually repaid, which turns out to be common, the *timing* of taxes yields an interest-free loan for which $0<e<1$.

The derivations required to completely characterize these dimensions of experience rating are contained in the Appendix. The results of this analysis may be summarized as follows. In a steady state, any employer's history of generating *UI* payments to workers is completely determined by his long-run equilibrium rate of insured unemployment, which I denote by $\mu$. If the employer's tax contributions per worker are to balance the flow of benefits generated by $\mu$, there must exist a tax rate, $\tau$, such that $\tau W=\mu B$, where $W$ is the taxable wage base per employee and $B$ is *UI* benefits expressed as an annual rate. Therefore, in order that $\mu$ be sustainable in this sense, $\tau$ must lie between the minimum and maximum rates charged in the employer's state. Letting $\rho=W/B$, this requirement can be expressed in terms of a pair of bounds on the equilibrium unemployment rate $\mu$:

$$(1) \qquad \rho\tau_{\min}\leqslant\mu\leqslant\rho\tau_{\max}.$$

If $\mu$ lies outside the range given in (1), the employer pays $\tau_{\min}$ or $\tau_{\max}$ independently of his layoff behavior, and so $e=0$. Therefore, the bounds defined by (1) determine the *range* of equilibrium unemployment rates that are subject to experience rating within state systems.[13] The relation of these bounds to the data in Table 2 on negative balance firms is obvious.

If $\mu$ lies within the range given in (1), tax dynamics in response to changes in an employer's layoff behavior determine the degree of experience rating. These dynamics are nontrivial, and they differ substantially across state systems, but their implications can be summarized by the shape of the cost function shown in Figure 1. In the illustrated case, which depicts experience rating in a typical state in 1975, rated firms paid about 80 cents in future taxes for each incremental dollar of benefits received by their workers. Nonrated firms with unemployment rates greater than $\mu_{\max}$ or smaller than $\mu_{\min}$ paid

---

[11] For a summary of these laws, see Becker (1972) and U.S. Department of Labor (1979).

[12] The wage base chosen may not fall below a federally mandated level, which now stands at $6,000.

[13] Equation (1) applies exactly in reserve-ratio systems, but it requires slight modification for benefit-ratio accounting. See the Appendix.

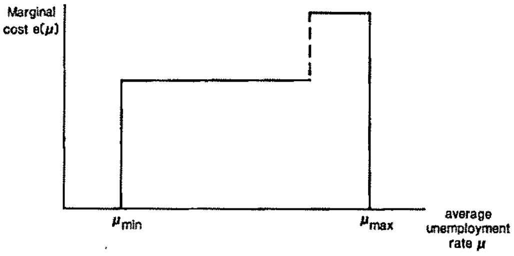


Figure 1. Marginal Cost of a Layoff Relative to Charged Benefits, Reserve-Ratio, and Benefit-Ratio States

Table 3—Experience Rating Parameters for Illustrative State Programs, 1975

| | Experience Rating Method | $\mu_{\min}$ | $e(\mu)$ | $\mu_{\max}$ |
|---|---|---|---|---|
| California | *RR* | 0.9 | .58 | 3.8 |
| Florida | *BR* | 0.0 | .99 | 6.2 |
| Massachusetts | *RR* | 3.2 | .90 | 4.6 |
| Michigan | *RR* | 0.1 | .86 | 7.0 |
| New York | *RR* | 1.5 | .71 | 5.3 |
| Ohio | *RR* | 0.2 | .75 | 4.0 |
| Texas | *BR* | 1.5 | 1.14 | 6.5 |
| Wisconsin | *RR* | 0.0 | .81 | 4.5 |

*Notes*: $\mu_{\min}$ and $\mu_{\max}$ are calculated using maximum and minimum tax rates in effect and the state ratio of average weekly benefits to average weekly taxable wages. The data are from *Handbook of Unemployment Insurance Financial Data* (1978), *Comparisons of State Unemployment Insurance Laws*, and unpublished U.S. Department of Labor data. *RR* denotes reserve ratio accounting, and *BR* denotes benefit ratio.

nothing. In some states, depending on the speed with which taxes are adjusted, this marginal cost to rated firms may actually exceed \$1 ($e > 1$) and, as shown, $e$ need not be constant throughout the rated range.

Estimates of the bounds in (1) that are relevant for different state systems may be calculated from information on maximum and minimum tax rates, and on the levels of benefits and taxable wages in each state. Estimates of the degree of experience rating in each state can be similarly derived from the formulae used by each state to calculate taxes.[14] Some values for both margins of experience rating are shown in Table 3 for illustrative state programs. These data show that both the range of unemployment over which experience rating is relevant and the degree of rating within that range differ substantially across state systems. Thus, for example, in 1975, experience rating for firms and their workers in Michigan was truncated above at an insured unemployment rate of 7 percent, while this upper bound occurred at only 3.8 percent in California. This means that for employers with equilibrium rates of, say, 6 percent, *UI* would increase the value of unemployment relative to employment in California by the full amount of available benefits, while 86 percent of *UI* would be repaid in Michigan. In terms of magnitudes of incentives, at a typical 1975 weekly benefit amount of \$90, that is a difference in costs of \$77 per week.[15] Even within the range where the marginal cost of benefits is positive, a California firm would repay only 58 cents of each dollar received by an additional unemployed worker. At the other extreme, an identically situated firm in Texas would repay *more* than the value of benefits that workers receive, so that layoffs are actually taxed for rated employers.

By any standard, these differences in layoff incentives offered to otherwise identical firms are large. It is this type of variation in the two relevant margins of experience rating —its range and degree—that may be exploited in the cross section to estimate the impact of these incentives on layoff decisions.[16] I turn to the empirical methodology for identifying these effects.

## II. The Data and an Empirical Strategy

### A. *The CPS Data and the Basic Model*

To estimate the impact of these differences in incentives on layoff unemployment, I utilize a sample of more than 8,000 individuals selected from the March 1975 file of the

---

[14] Described in the Appendix.

[15] These calculations ignore the additional effect on the subsidy caused by the tax-exempt nature of most *UI* payments. This omission is corrected in the econometric analysis.

[16] For purposes of empirical analysis, I treat these differences in costs as exogenous and do not speculate on the political forces that may support them.

*Current Population Survey*.[17] Simply stated, the empirical strategy links to each observation in the *CPS* information on the level of available *UI* benefits and the characteristics of experience rating that apply in an individual's state and industry. Using straightforward econometric techniques, it is then possible to decompose the effect of these variables on the probability that an individual is on layoff into component probabilities of transiting to and from layoff unemployment. The particular virtue of the March *CPS* file for this purpose is its wealth of information on the personal characteristics and employment status of individuals, which includes retrospective information on earnings and employment during the calendar year that preceded the survey. This retrospective data is essential, since it is used to establish the level of *UI* benefits for which each individual qualifies. The relative incidence of temporary layoff unemployment was fairly high in 1975, which facilitates an analysis of the factors that influence layoff and rehire decisions. Hence my choice of that year.

To minimize the influence of vagaries in individual labor force participation decisions on the results, the sample includes only persons who were between the ages of 20 and 65, and who were full-time, full-year labor force participants during the previous year. In addition, to focus on persons who were subjective to layoff risks, I selected privately employed, nonprofessional workers who were not self-employed and who did not work in the agricultural sector.[18] A final selection criterion was imposed by the data: the *CPS* reports "state" of residence for each observation, but smaller states are aggregated into larger geographic units to protect the confidentiality of responses. Since the identification of *UI* effects depends crucially on differences among states in methods of financing and liberality of benefits, I consider only individuals for whom state of residence could be exactly identified. The resulting sample of 8,280 individuals represented 19 reserve-ratio and benefit-ratio financing systems and 29 two-digit (SIC) industry classifications.[19] Of these, 555 individuals (6.7 percent of the sample) were on temporary layoff. Since the economywide layoff rate was slightly above 3 percent in this year, it is clear that this sample faces greater layoff risk than the general population. No attempt is made here to extend the results to other, more heterogeneous populations or to those with weaker labor force attachment.

The estimation strategy uses two distinct pieces of information on an individual's labor market status to estimate the effects of *UI*, and other variables, on transitions to and from layoff unemployment: current status (for example, on temporary layoff) and, for unemployed workers, the duration of the current spell in progress. To use this information, certain assumptions on the processes that generate transitions are required. I assume that the underlying processes are stationary, and that the per period probabilities of entering and leaving layoff unemployment are constant, independent of time spent employed or unemployed. Denote these constant "hazards" by $\lambda^{el}$ and $\lambda^{le}$ for transitions from employment to layoff, and from layoff to employment, respectively.[20] With this structure, the Appendix establishes that in a sample of individuals who are employed or on layoff, the contribution to the sample likelihood by individual $j$ who is on layoff at the survey date is

$$(2) \qquad L_j = \frac{\lambda_j^{el}\lambda_j^{le}\exp\left[-T_j\lambda_j^{le}\right]}{\lambda_j^{le}+\lambda_j^{el}},$$

---

[17] The *CPS* is a monthly survey from which the government's employment and unemployment statistics are compiled. The survey is slightly different in each month, and only the March file contains information on employment and earnings during the previous year.

[18] A complete listing of selection criteria and sample characteristics may be obtained from the author.

[19] These included all manufacturing industries, plus construction, mining, wholesale trade, retail trade, and transportation industries.

[20] The stationarity and constant hazards assumptions imply that transitions are first-order Markov. Thus, the probability that a person is on layoff at the sample date is simply the proportion of time that an infinitely lived person in such an environment would spend unemployed. Formally, the hazard is the conditional probability of passage given duration $T$. Constancy implies that this probability is independent of $T$, and so the distribution of completed spell lengths is negative exponential. Further details are appended.

where $T_j$ is the duration of the layoff spell in progress. Equation (2) is simply the joint probability of observing an individual on layoff *and* in the $T$th week of his spell. In contrast, if the individual is employed then duration in employment is not observed, and so the contribution to the likelihood is simply

$$L_j = \lambda_j^{le} \Big/ \left[\lambda_j^{el} + \lambda_j^{le}\right]. \tag{3}$$

The transition rates $\lambda^{el}$ and $\lambda^{le}$ are the basic elements that we seek to identify in the estimation procedure. Intuitively, under the stated assumptions the probabilities that a randomly selected individual is employed or on layoff depend on *both* these transition rates, as in (3). The single piece of information given by the empirical frequency of layoff spells in the sample would then be inadequate to identify each of $\lambda^{el}$ and $\lambda^{le}$.[21] However, the information given by the empirical distribution of durations, $T$, among unemployed workers allows us to estimate $\lambda^{le}$—the rate at which individuals leave unemployment—separately. This estimate may then be used, via (3), to identify $\lambda^{el}$. Estimating the two transition rates jointly from the full sample of individuals who are both employed and on layoff enhances efficiency.

To close the empirical model, I specify the dependence of $\lambda^{le}$ on a vector of explanatory variables, $X_j$, for the $j$th individual as

$$\lambda_j^{le} = \exp\{\mathbf{X}_j \beta^{le}\}, \tag{4}$$

and $\lambda^{el}$, the transition rate from employment, is written conformably. Equation (4) is nonnegative, as required, for all values of the vectors $\beta$, which are to be estimated. The number of individual parameters to be estimated is twice the number of regressors in **X**. Thus, for example, $\beta^{el}$ measures the impact of exogenous variables on the probability of *entering* unemployment, while (3) implies that the total impact of **X** on the probability of unemployment (the unemployment rate) is measured by the difference $\beta^{el} - \beta^{le}$.[22] The likelihood defined by equations (2), (3), and (4) is globally concave, and so standard techniques are effective in maximizing it. Summary statistics for the variables included in **X** are reported in Appendix Table A1, and are discussed below.

### B. *Specifying and Identifying the Effects of UI*

In estimating this model, I focus on two *UI* variables as determinants of transitions, and thus of employment status at a point in time. These are the benefit-replacement ratio and the amount of benefits that are subsidized by methods of financing. The former variable is common in empirical work on the effects of *UI*, and in this specification it captures incentive effects that do not depend on imperfections in experience rating or the *UI* subsidy. Potential benefits were imputed for each observation from available personal information and from the qualifying provisions of each state's *UI* law,[23] while the relevant marginal tax rates were calculated from state and federal tax tables based on family structure and on earnings during the previous calendar year. As Table A1 indicates, for the typical individual in this sample, available *UI* benefits would have replaced 56 percent of spendable weekly earnings during a spell of unemployment. Sources of variance in this variable both within and between state programs were illustrated in Table 1 above.

Imputation of the *UI* subsidy that is relevant for each individual is based directly on the cost functions described above. The subsidy to unemployment occurs both because of imperfect experience rating and because *UI* benefits are not normally subject to income taxation. Thus, even with full experience rating ($e(\mu)=1$), unemployment would be subsidized by the different tax treatments of earnings and benefits. Combining these two incentives, the change in the value of unemployment induced by a weekly benefit amount of \$$b$ is simply

---

[21] With knowledge of only the relative frequency of spells, the ratio $\lambda^{le}/\lambda^{el}$ could be identified from (3).

[22] Equations (3) and (4) imply that the unemployment probability is determined by a logistic distribution. Thus $\beta = \beta^{el} - \beta^{le}$ is the coefficient on $x$ in a standard logit model of unemployment incidence.

[23] The required information included weeks worked during the previous year, weekly earnings, and family structure (to calculate dependent's allowances). Given heterogeneity in state laws, the computer program for this imputation is fairly complex. It is available on request from the author.

$$b((1/1-t)-e(\mu)), \tag{5}$$

where $t$ is the individual's marginal tax rate on earned income.[24] The crucial element of (5) is $e(\mu)$, so that precise imputation of the subsidy requires information on $\mu$ for each individual's employer. Such detailed information on unemployment histories is clearly unavailable. In its place, I utilize time-series data on insured unemployment rates at the two-digit SIC level of aggregation,[25] and corresponding *CPS* information on the industry classification of each observation's most recent employer. I assume that in each industry the equilibrium insured unemployment rate is equal to the actual insured rate over the period 1971–74. Calling this rate in industry $i$, $\bar{\mu}_i$, the simplest imputation method would assign $e_s(\bar{\mu}_i)$ for individuals in state $s$. With slight modification to account for within-industry variation in equilibrium rates among employers, this method was used to impute the degree of experience rating in the 551 state-industry cells.

The modification simply recognizes that, within any industry, equilibrium rates will be distributed about the industry mean. Consequently, even in industries where $\bar{\mu}_i$ lies outside a state's experience rated range defined in (1), experience rating may affect layoff decisions for some firms and conversely for industries where $\bar{\mu}_i$ is within the rated interval. There is no information on the form of these distributions, and so any allowance for this type of effect must be somewhat crude. Thus, since the potential for within-industry variance is larger in industries with high average unemployment, I assume an industry density for $\mu$, $f_i(\mu)$, that is triangular with the range of variation in $\mu$ proportional to the mean.[26] The factor of proportionality $\gamma$ is the same in all industries, and so the assumed range of $\mu$ is equal to $\bar{\mu}_i \pm \gamma\bar{\mu}_i$ in the $i$th industry. The *expected* degree of experience rating for person $j$ in industry $i$ may then be imputed using the density $f_i(\mu)$ and the cost function $e(\mu)$ in each state program. The parameter $\gamma$ may be chosen on the basis of the model's overall fit, and a value of $\gamma > 0$ indicates effects of experience rating even where $\bar{\mu}_i > \mu_{\max}$. Of course, because of the form of $f(\mu)$, this effect is constrained to die out the farther is $\bar{\mu}_i$ from the experience rated range.

The fact remains that industries with extremely high or low average unemployment will have less imputed experience rating—since a larger proportion of $f_i(\mu)$ would lie outside the range defined in (1)—and this may seriously bias the results in favor of powerful experience rating effects.[27] This source of bias may be purged by estimating the model with vectors of fixed industry effects in the specifications of both $\lambda^{el}$ and $\lambda^{le}$. That is, I assume

$$\mathbf{x}_j\beta^{el} = \alpha_i^{el} + \mathbf{Z}_j\theta^{el}, \tag{6}$$

and $$\mathbf{x}_j\beta^{le} = \alpha_i^{le} + \mathbf{Z}_j\theta^{le},$$

---

[24] Currently, workers who earn more than \$18,000 annually are subject to federal income taxation on their *UI* benefits. This is a recent development, however, and in 1975 all benefits were tax exempt.

[25] These are nationwide unemployment rates for each industry. They appear in various issues of *Unemployment Insurance Statistics*. Some series were unpublished, but they may be obtained from the Department of Labor.

[26] The density is of the form

$$f_i(\mu) = \begin{cases} (1+(\mu-\bar{\mu}_i))/(\gamma\bar{\mu}_i)^2 & \mu < \bar{\mu}_i; \\ (1-(\mu-\mu_i))/(\gamma\mu_i)^2 & \mu > \mu_i. \end{cases}$$

[27] Table 5, below, makes this especially clear. Industries such as Apparel and Miscellaneous Manufacturing have relatively high values of $\bar{\mu}_i$, high sample layoff rates, and large imputed subsidies. Even if the subsidy did not affect layoffs, this relationship would almost surely yield a positive estimated effect of the subsidy on unemployment. The strategy indicated in the text simply subtracts these cross-industry effects, and concentrates solely on variation in the subsidy within industries.

where the $\alpha_i$ represent industry-specific shifters of the transition rates.[28] In this specification, differences in levels of exogenous variables across industries are captured by the $\alpha_i$, so only *within-industry* variation in $\mathbf{Z}_j$ is used in estimating $\theta^{el}$ and $\theta^{le}$.

Since $\bar{\mu}_i$ is the same for all persons in industry $i$, within-industry variation in the degree of experience rating is caused only by differences between states in the location and height of the cost function, $e(\mu)$. Therefore, any estimated impact of the *UI* subsidy reflects only the within-industry experiment of changing the function $e(\mu)$, while differences between industries in average levels of the subsidy and unemployment do not affect the results.

## III. The Evidence

All of the specifications that I have estimated show strong effects of unemployment insurance on layoff unemployment. More importantly, the evidence is that most of *UI*'s impact on layoffs is caused by current methods of experience rating. In this sample, the layoff unemployment rate would have been about 30 percent lower if the subsidy to unemployment caused by the current *UI* system had been eliminated. In contrast, simple changes in the level of *UI* benefits that leave the subsidy constant would have only minor effects on layoffs.

Table 4 summarizes the main results for the determinants of temporary layoff unemployment. In the table, I report two specifications of the model, the first of which ignores any information on experience rating. Thus, all *UI* effects in this specification are captured by the replacement ratio, and the impact of this variable will reflect the average degree of *UI* subsidization in addition to any independent effects of changing the level of available *UI*. The estimates in the first specification show a strong positive impact of *UI* on unemployment, and the dominant share of this effect is due to a statistically significant increase in the probability of experiencing a layoff, rather than to increased duration of spells. To fix ideas on the magnitudes of these effects, evaluating the model at sample means yields an estimated monthly probability of entering unemployment of .029. At means, the point estimate of the ratio's impact on this transition implies that the per period layoff probability is unit elastic with respect to benefits $(1.82 \times .56 = 1.02)$, so a 10 percent across-the-board reduction in the level of *UI* (about \$8 per week) would reduce the monthly probability of entering layoff by about 0.3 percentage points. At the other margin, however, there is no significant effect of the replacement ratio on the probability of leaving layoff.

Overall, the point estimate of *UI*'s total effect in the third column (1.70) implies that a 10 percent reduction in the level of benefits would eliminate about 8.8 percent of all layoff spells in this sample.[29] I have experimented with various nonlinear forms for the effects of the ratio (for example, splines and quadratics), but these findings were not affected by such changes in specification. In short, the estimates in the first specification indicate that the incidence of temporary layoff unemployment is quite sensitive to the availability of unemployment insurance, and that transitions *into* layoff unemployment play a prominent role in explaining this relationship.

The effects of other exogenous variables in model A are largely self-explanatory. Recall that the specification controls for differences caused by the characteristics of particular industries, so the estimates reflect within-industry changes in exogenous variables. For example, since the usual statistical determinants of earnings are controlled for in the

---

[28] Since there are 29 industries represented in the data, the model estimates 58 industry coefficients in addition to the other parameters of the model. In practice, this brought to 78 the number of parameters in the estimated models. The procedure for estimating the fixed effects in a nonlinear model like this one is considerably more complicated than simply subtracting out industry means and it is expensive to compute.

[29] This total impact of *UI* is similar to that of Feldstein (1978), who regressed layoff unemployment on the replacement ratio. Feldstein's data were for a different year (1971) and a more heterogeneous sample, which makes comparisons difficult. Note that Feldstein's *OLS* method of estimating the probability of a layoff is not capable of identifying separate effects on transitions.

TABLE 4—UNEMPLOYMENT INSURANCE AND LAYOFF UNEMPLOYMENT

| | Specification A | | | | Specification B | | | |
|---|---|---|---|---|---|---|---|---|
| Exogenous Variables: $Z_j$ | Entering Layoff[a] | Leaving Layoff[b] | Total Effect[c] | Marginal Probability[d] | Entering Layoff | Leaving Layoff | Total Effect | Marginal Probability |
| Subsidy | | | | | .683 | −.402 | 1.09 | .072 |
| | | | | | (1.35) | (−1.05) | (3.29) | |
| Replacement Ratio | 1.82 | .117 | 1.70 | .112 | 1.01 | .414 | .603 | .040 |
| | (2.66) | (.241) | (3.54) | | (1.24) | (.729) | (1.02) | |
| Weekly Wage ÷ 100 | .047 | .065 | −.018 | −.001 | −.023 | .061 | −.084 | −.005 |
| | (.404) | (.794) | (−.217) | | (−.193) | (.730) | (−.967) | |
| Age | −.058 | .034 | −.092 | −.006 | −.058 | .036 | −.094 | −.006 |
| | (−1.52) | (1.27) | (−3.38) | | (−1.50) | (1.29) | (−3.45) | |
| $Age^2 \div 100$ | .036 | −.047 | .083 | .005 | .036 | −.048 | .084 | .005 |
| | (.761) | (−1.39) | (2.51) | | (.752) | (−1.39) | (2.49) | |
| Age = 25[e] | −.040 | .011 | −.051 | −.003 | −.040 | .011 | −.051 | −.003 |
| | (−2.67) | (1.03) | (−4.85) | | (−2.63) | (1.01) | (−4.82) | |
| Age = 40[e] | −.029 | −.003 | −.026 | −.002 | −.029 | −.003 | −.026 | −.002 |
| | (−5.25) | (−.829) | (−6.23) | | (−5.20) | (−.819) | (−6.20) | |
| Education | −.050 | −.068 | .018 | .001 | −.046 | −.072 | .026 | −.002 |
| | (−.428) | (−.890) | (.204) | | (−.389) | (−.930) | (.299) | |
| $Education^2 \div 100$ | −.066 | .405 | −.471 | −.031 | −.086 | .432 | −.519 | −.034 |
| | (−.119) | (1.08) | (−.732) | | (−.154) | (1.15) | (−1.24) | |
| Education = 12[e] | −.066 | .027 | −.093 | −.006 | −.066 | .031 | −.097 | −.006 |
| | (−1.96) | (1.15) | (−3.88) | | (−1.96) | (1.32) | (4.05) | |
| Race (White) | −.092 | .044 | −.136 | −.009 | −.107 | .050 | −.157 | .010 |
| | (−4.91) | (.388) | (−1.00) | | (−.563) | (.385) | (−1.14) | |
| Sex (Male) | .177 | .224 | −.047 | −.003 | .199 | .226 | −.027 | −.002 |
| | (1.00) | (1.79) | (−3.75) | | (1.22) | (1.79) | (−.219) | |
| Industry Effects | Yes | Yes | Yes | | Yes | Yes | Yes | |
| $\ln \mathcal{L}$ | −3636.5 | | | | −3631.2 | | | |
| Proportion Unemployed | .071 | | | | | | | |

*Note*: Asymptotic normal statistics in parentheses.

[a] Estimate of $\theta^{el}$.

[b] Estimate of $\theta^{le}$.

[c] Estimate of $\theta^{el} - \theta^{le}$.

[d] Derivative of unemployment probability at sample mean.

[e] Age and education quadratics evaluated at the indicated level.

specification, the coefficient on the weekly wage might be interpreted as the effect of unobserved components of individual or job-specific productivity. There is no evidence in these data that unemployed workers are materially different in this respect than their employed counterparts, nor are the individual transition rates significantly affected by earnings capacity. This is not too surprising, since most workers on layoff will end up returning to their old jobs. The effects of age and education, however, are strong. On average, a 25-year old is about 50 percent more likely to be unemployed than a 40-year old (wage constant), and an extra year of schooling reduces unemployment by over half a point. The age profile of unemployment probabilities is nonlinear, becoming flatter at older ages. As with the impact of the replacement ratio, these effects are concentrated on the probability of entering layoff.[30] Finally, there are no significant effects of race or sex on layoff unemployment rates, although men have slightly shorter spells than women. These results are not sensitive to changes in the specification of *UI* effects, and so they will not be referred to further below.[31]

The imputed *UI* subsidy to unemployment is controlled for in specification B. It turned out that the dispersion parameter $\gamma$ was quite costly to estimate directly in the maximum likelihood procedure, though in experiments with the imputation the data preferred a value of $\gamma = 0.25$. This value implies a within-industry standard deviation in $\mu$ of only about 15 percent of the mean, and the results are not statistically different from those obtained when $\gamma = 0$. Thus, there is little evidence that within-industry dispersion of $\mu$ is an important consideration in evaluating the incentive effects of experience rating. The direct impact of imputed experience rating on layoff unemployment, however, is both statistically and numerically significant.

In the specification, I have entered the subsidy as a proportion of weekly earnings so that the imputed degree of rating, $e(\mu)$, is the only source of independent variation in the subsidy relative to the replacement ratio.[32] In light of the powerful effect of the replacement ratio estimated in the first model, this restriction offers a strong test of the role of experience rating in affecting layoff decisions, and serves to divide the total impact of *UI* between its subsidized and nonsubsidized components. Thus, note that when the subsidy is controlled for the estimated impact of the replacement ratio falls by over 60 percent relative to model A, and it is no longer significant by standard criteria. At means, the new point estimate of .603 implies that a 10 percent change in the benefit level that holds the subsidy constant would change the layoff probability by two-tenths of a percentage point.[33] In contrast, the *UI* subsidy has a powerful and statistically significant effect on the layoff unemployment rate. Since the average subsidy is equal to 31 percent of the weekly wage, the point estimate in the last column implies that the subsidy accounts for (approximately) $.072 \times .31/.071 = 31$ percent of the sample's layoffs. In other words, if the incentives toward increased unemployment offered by current methods of experience rating and income taxation were entirely eliminated, the average layoff unemployment rate in this sample would decline by more than one fourth.

The data and discussion of Section I imply that current methods of financing *UI* will have a larger impact on layoffs in high unemployment industries, where the subsidy to unemployment is larger and experience rating less relevant to layoff decisions. Table 5 illustrates this distributional effect, showing the predicted amount of the layoff unemployment rate that is accounted for by the subsidy in selected industries. Thus, for example, in the high unemployment Apparel industry ($\bar{\mu} = 7.65$ percent), the subsidy is equal to about 56 percent of weekly earnings, on average, and so it accounts for over 6.7 points of the industry layoff rate in this year. In contrast, in Primary Metals, average un-

---

[30] This finding is consistent with that of Robert Hall (1972) and others, who found that differences in the duration of unemployment across groups are of minor significance in explaining differences in unemployment rates. That this also holds for temporary layoffs is somewhat more surprising, given the symmetry of layoff and rehire decisions.

[31] Note that the worker's union status is not an explanatory variable. This is because union membership is not reported in the March file (except in 1971). However, any differences across industries in unionization rates are captured by the industry effects, $\alpha_i$. It is therefore highly doubtful that the omission of union status could strongly affect the results.

[32] Since the replacement ratio is $b/w(1-t)$, the subsidy as a proportion of weekly earnings multiplies the ratio by the quantity $1-e(\mu)(1-t)$. This is an interaction term, and the results imply that most of the effect of *UI* comes through this interaction with experience rating.

[33] The point estimate of 1.01 in the transition to layoff is still numerically large, but it is imprecisely estimated and is partially offset by a *positive* and insignificant estimated effect on transitions from layoff. Note that at the mean of the subsidy (.31), the effect of the same 10 percent change in benefits is also about .2 points ($.072 \times .31 \times .1 = .0022$), so that the total effect is almost evenly divided. When local market conditions are controlled for in Table 6, however, the share of this effect shifts strongly to the subsidy.

TABLE 5—ESTIMATED EFFECTS OF THE UNEMPLOYMENT INSURANCE SUBSIDY, BY INDUSTRY, 1975

| | $\bar{\mu}_i$[a] | Subsidy/ Weekly Wage | Layoff Unemployment Rate[b] | Total Effect of Subsidy (percent)[c] |
|---|---|---|---|---|
| Wholesale Trade | 1.82 | .33 | 1.86 | 0.65 |
| Machinery | 1.92 | .32 | 4.44 | 1.47 |
| Chemicals | 1.96 | .32 | 1.78 | 0.60 |
| Retail Trade | 1.99 | .27 | 1.98 | 0.58 |
| Primary Metals | 2.43 | .23 | 7.34 | 1.70 |
| Electrical Machinery | 3.10 | .29 | 8.38 | 2.46 |
| Fabricated Metals | 3.31 | .27 | 11.72 | 3.04 |
| Food | 4.73 | .36 | 5.10 | 1.88 |
| Misc. Manufacturing | 6.77 | .61 | 9.82 | 5.86 |
| Apparel | 7.65 | .57 | 12.54 | 6.76 |
| Construction | 8.33 | .42 | 14.15 | 5.56 |
| Total | 3.75 | .31 | 7.1 | 2.22 |

[a] Calculated from average monthly *UI* compensated unemployment rates, by industry, 1972–74. *Source*: U.S. Department of Labor, *Handbook of Unemployment Insurance Financial Data*, various issues.

[b] Workers on layoff as a proportion of workers employed and on layoff.

[c] Approximate effect of eliminating the imputed *UI* subsidy. The formula is total effect $= 1.09 \times (\bar{p}_i) \times (1 - \bar{p}_i) \times$subsidy, using model (2)–(4), where $\bar{p}_i$ is the industry layoff rate.

employment is much lower ($\bar{\mu} = 2.43$ percent) and experience rating more extensive, so the mean imputed subsidy accounts for only 1.7 points of total layoff unemployment.

The decomposition of the total impact of the subsidy into its component effects on probabilities of entering and leaving layoff also shows the expected signs, though the point estimates are only moderately larger than their standard errors. Using the point estimates, 62 percent of the total effect on unemployment is due to increased transitions from employment to layoff (the layoff probability is 20 percent higher), while the point estimate of $-.402$ in layoff transitions implies that the mean level of the subsidy increases the average duration of layoffs by about 1.3 weeks. Thus, though the estimates from the decomposition are imprecise, there is evidence in these data that the significant impact of the subsidy on layoff unemployment is due to both more frequent and longer layoff spells. And as in Table 5, these effects will be nonneutral across industries, being larger in poorly experience rated sectors.

Changes in the reported specification that excluded various combinations of demographic characteristics did not materially affect the results, nor did restrictions of the sample to men only, or to white men only. However, one change in the basic specification does affect the estimated impact of unemployment insurance. Recall that the results shown in Table 4 rely heavily on between-state differences in the liberality of benefits, and the extent of experience rating to estimate incentive effects of *UI*. Thus, to the extent that there are important differences across states in "local" market conditions affecting layoff, rehire, and search decisions, these results may be biased.[34] For example, rapidly growing states such as Texas and Florida are reputed to be "strong" local markets, and the data of Tables 1 and 3 reveal relatively low benefit levels and strong experience rating in those states. To control for these long-term differences among areas, I have reestimated the model in Table 5 including as an explanatory variable the imputed growth rates of private, nonagricultural employment in each individual's state. These rates were estimated from quadratic trend regressions for each state, and the re-

---

[34] I am indebted to Bob King for raising this issue.

TABLE 6—ESTIMATED EFFECTS OF UNEMPLOYMENT INSURANCE AND STATE GROWTH OF LAYOFF UNEMPLOYMENT

| | Specification A | | | Specification B | | |
|---|---|---|---|---|---|---|
| Exogenous Variables: $Z_j$ | Entering Layoff[a] | Leaving Layoff[b] | Total Effect[c] | Entering Layoff[a] | Leaving Layoff[b] | Total Effect[c] |
| Subsidy | .612<br>(1.21) | −.396<br>(−1.04) | 1.007<br>(3.01) | .874<br>(1.98) | −.219<br>(.649) | 1.093<br>(3.82) |
| Replacement Ratio | .861<br>(1.01) | .559<br>(.949) | .302<br>(.491) | — | — | — |
| Growth Rate | −.426<br>(−1.18) | .205<br>(.874) | −.631<br>(−2.28) | −.519<br>(−1.50) | .141<br>(0.32) | −.660<br>(−2.45) |
| $\ln \mathcal{L}$ | −3926.2 | | | −3627.9 | | |

*Note*: Asymptotic normal statistics in parentheses. Other explanatory variables are as in Table 4.

[a] Estimate of $\theta^{el}$.

[b] Estimate of $\theta^{le}$.

[c] Estimate of $\theta^{el}-\theta^{le}$.

sults of including them are summarized in Table 6.[35]

Even after controlling for industry composition, rapidly growing states have fewer layoffs, and the majority of this effect is again attributable to the probability of entering unemployment. At means, a one percentage point increase in long-term growth (less than a standard deviation) reduces the layoff rate by about .5 points. Most importantly, however, controlling for local growth reduces the impact of the replacement ratio by about 50 percent from the estimate in Table 4. In fact, the impact of the ratio is now negligible: a 10 percent change in the benefit level that leaves the subsidy constant would change the layoff unemployment rate by only one-tenth of a point, though the standard error on this estimate is large.[36] The effects of the unemployment insurance subsidy, however, are not materially different than in Table 4, and its total effect on unemployment remains highly significant. For completeness, I also report in the table the estimated effect of the *UI* subsidy when the replacement ratio is excluded. Again, the results are not highly sensitive to this exclusion, though the impact on transitions into unemployment is somewhat larger in this case. I conclude from this evidence that the impact of *UI* is overstated when local market conditions are ignored, but this bias falls most on the replacement ratio. Methods of financing *UI*, in contrast, continue to have strong and stable effects on layoffs that are consistent with theory.

## IV. Summary and Concluding Remarks

The vast majority of employer-initiated spells of unemployment involve the payment of unemployment insurance benefits. This empirical connection between layoff decisions and the *UI* system may affect observed unemployment rates for two reasons. First, the legislated level of benefits that must be paid to laid off workers can change the relative costs of unemployment to both workers and their employers, and this alone may influence layoff, rehire, and search decisions. Second—and this has been the main focus of this paper—current systems of *UI* financing clearly subsidize the occurrence of unemployment, since most employers are only partially liable for the benefits that their workers receive. Largely for lack of suitable measures of the incentives offered to firms, however, these effects of unemployment insurance on the decisions that generate unemployment have been previously untested.

---

[35] The data for these auxiliary regressions are taken from the U.S. Department of Labor, *Employment and Earnings, States and Areas*, various issues.

[36] The total effect of a 10 percent change in benefits that *includes* the effect of experience rating is, at means, .32 percentage points. Two thirds of this effect is accounted for by the subsidy. This total effect can also be decomposed into effects on entering and leaving unemployment, and it is clear from the point estimates that passages to layoff unemployment are more strongly affected by *UI*.

The empirical evidence presented in this paper indicates that both the extent and the effect of *UI* subsidization are important. Examination of the structures of state financing laws shows that even experience-rated firms often face significant unemployment subsidies, and an important portion of aggregate employment is not subject to experience rating at all. The econometric analysis showed that the impact of the unemployment insurance subsidy on layoff unemployment is powerful—the imputed subsidy accounts for more than a quarter of all layoffs in the data—and the effects on the probabilities of entering and leaving unemployment are in accord with theory. For workers who are subject to temporary layoff, *UI* appears to have a more important effect on transitions *into* unemployment than on the more commonly studied duration of spells. Taken together, these findings offer strong support for the hypothesis, originally proposed by Feldstein (1973), that methods of financing unemployment insurance can have important effects on the incidence of unemployment.

These results imply that, *without changing benefit levels available to unemployed workers*, a significant reduction in layoff unemployment could be achieved by changing the incentives offered by current *UI* laws. It is tempting to conclude from these findings that subsidies to unemployment should be eliminated via complete experience rating of *UI* taxes (full employer liability) and the symmetric tax treatment of benefits and earned income. My analysis does not justify that conclusion, however, since very little is known about the optimal structure of *UI* financing systems. For example, under circumstances where labor turnover may be viewed as a public good (see Hall, 1979, and Peter Diamond, 1981), firms (and workers) will undervalue separations and so complete experience rating of *UI* may inefficiently discourage permanent layoffs. These arguments are far less applicable for temporary layoffs, however, and they cannot be used to justify the current structure of experience rating where many layoff-prone employers face no marginal cost of benefits at all. On these grounds, the empirical effects identified in this paper can be viewed as costs of the current organization of unemployment insurance financing in the United States.

## APPENDIX

### A. *Calculating the Degree of Experience Rating*

1. *Benefit-Ratio Systems*. In benefit-ratio systems, an employer's tax rate depends on the ratio of total benefits charged to the employer's account over the past $T$ years to total taxable wages for the same period. This is the benefit ratio:

$$BR = \sum_{j=1}^{T} B\mu_{t-j}N_{t-j} \Big/ \sum_{j=1}^{T} WN_{t-j},$$

or $$BR = \rho \sum_{j=1}^{T} n_{t-j}\mu_{t-j},$$

where $\rho = B/W$ and $n_{t-j} = N_{t-j}/\Sigma N_{t-k}$ is the share of year $t-j$ employment in total employment over the past $T$ years. Thus, the benefit ratio is just a share-weighted average of past unemployment rates times the "charge rate," $\rho$. In some benefit-ratio states, the firm's tax rate is just $\tau_t = BR_t$, but, in others, $BR_t$ is multiplied by a factor of proportionality, $\lambda$, equal to the ratio of total state benefit payments to those which are charged to firm accounts. Thus $\tau_t = \lambda BR_t$ with $\lambda \geqslant 1$. The present value of taxes caused by a transitory change in $\mu$ relative to benefits received may then be calculated to be

$$e = \lambda\left(1-(1+i)^{-T}\right)/T_i,$$

where $i$ is the rate of interest. This marginal cost of benefits is less than unity (for $\lambda = 1$) and is declining in $T$ because the implicit interest-free loan is repaid over a longer period when $T$ rises. In most states, $T = 3$ years, while the currently operating system in Michigan sets $T = 5$ years. Thus, with $T = 3$ and $i = .1$, $e = .828\lambda$ while with $T = 5 e = .758\lambda$.

2. *Reserve-Ratio Systems.* Under this method of accounting, each employer's tax rate, $\tau$, depends on the ratio of total funds in its account to its total taxable payroll—the reserve ratio. If $R_t$ is total reserves credited to the employer's account in year $t$, $W$ the taxable wage base per employee, and $N$ the total number of employees, then the reserve ratio is $r_t = R_t/WN$. Assuming $WN_t$ is approximately constant (most state programs use weighted averages over several years), $r_t$ follows $r_t = r_{t-1} + \tau_t - \rho\mu_t$, where $\rho = B/W$.

Between the maximum and minimum rates, $\tau_t$ is defined as a step function of $r_t$. When these steps are small, the function may be treated as approximately linear: $\tau = \eta_0 - \eta_1 r$. Then taxes follow the difference equation

$$\tau_{t+1}W = \bar{\tau}_{t+1} = (1-\eta_1)\bar{\tau}_t + \eta_1 B\mu_t.$$

Using this equation, a current increment to $\mu_t$ generates future taxes worth $B\eta_1/(\eta_1 + i)$ where $i$ is the rate of interest. Thus, dividing by the value of benefits received, $e = \eta_1/(\eta_1 + i)$. In a typical system, $\eta_1 = .3$ so with $i = .1$, $e = .75$.

For a large step in the tax function at ratio $\hat{r}$, the linear approximation is less appropriate. In general, the step may be characterized as $\tau = \tau_1$ for $r < \hat{r}$ and $\tau = \tau_0$ for $r > \hat{r}$. The tax rate that would support a steady-state value of $\mu$ is $\tau^* = \rho\mu = \tau_0 + \phi(\tau_1 - \tau_0)$ where $0 < \phi < 1$. Thus, when $\tau = \tau_0$, reserves decline at rate

$$r_t - r_{t-1} = \tau_0 - \rho\mu_t = -\phi(\tau_1 - \tau_0) = -\phi\Delta\tau.$$

Conversely, if $\tau = \tau_1$, reserves accumulate at rate

$$r_t - r_{t-1} = \tau_1 - \rho\mu_t = (1-\phi)\Delta\tau.$$

Now the tax authority sets $\tau$ on the basis of $r$ at annual evaluations. A firm with $\tau_0 < \tau^* < \tau_1$ will find that its taxes alternate between $\tau_0$ and $\tau_1$ as its reserve ratio is above or below $\hat{r}$ at the evaluation dates. Define the beginning of such a cycle, $\hat{t}$, as the first instant where $r_t$ crosses $\hat{r}$ from above, and let $\alpha$ be the proportion of a year remaining from $\hat{t}$ to an evaluation. Assuming $\phi < .5$ (derivations for $\phi > .5$ are symmetrical), reserves decline to $\hat{r} - \Delta\tau\phi\alpha$ at the first evaluation, they rise for one year to $\hat{r} + \Delta\tau(1 - \phi(1+\alpha))$, and then decline to $\hat{r}$ after $\phi^{-1} - (1+\alpha)$ periods. Thus, the length of a cycle is $\phi^{-1}$, of which $\phi^{-1} - 1$ periods have a tax of $\tau_0$ and one has a tax of $\tau_1$. As of the first evaluation, the present value of the surcharge $\Delta\tau$ for one period is $\Delta\tau(1 - e^{-i})/i$. Assuming that $\alpha$ is uniformly distributed on the unit interval, the expected present value of the tax surcharge as of the period $\hat{t}$ is

$$\int_0^1 \Delta\tau \frac{(1-e^{-i})}{i} e^{-i\alpha}\, d\alpha = \Delta\tau\left(\frac{1-e^{-i}}{i}\right)^2.$$

Since this surcharge occurs every $\phi^{-1}$ periods, the present value of the firm's future tax rate is

$$\frac{\tau_0}{i} + \Delta\tau\left(\frac{1-e^{-i}}{i}\right)^2 + \Delta\tau\left(\frac{1-e^{-i}}{i}\right)\frac{e^{-i/\phi}}{1-e^{-i/\phi}}$$

Now consider a change in unemployment that lasts exactly one period and generates change in reserves $dr = -\rho\, d\mu$. This change implies that the end of the cycle occurs $dr/\phi\Delta\tau$ periods sooner, and so the change in the present value of taxes as measured from the point where the cycle would have ended is

$$\beta = \Delta\tau\left(\frac{1-e^{-i}}{i}\right)^2\left(\frac{(\exp(i\, dr/\phi\Delta\tau) - 1)}{1 - e^{i\phi^{-1}}}\right).$$

Discounting this value to the period $t_0$, where the shock commences, and assuming that $t_0$ is uniformly distributed on $(0, \phi^{-1})$, that is, that layoffs can begin at any time, yields an expected present value of the tax increment of

$$\text{(A1)}\qquad \beta\int_0^{\phi-1} \phi e^{-i(\phi^{-1}-t_0)}\, dt_0$$

$$= \left(\frac{1-e^{-i}}{i}\right)^2 (\exp(i\, dr/\phi\Delta\tau) - 1)\frac{\phi\Delta\tau}{i}.$$

The change in benefits received per unit of taxable payroll is just $dr$, so dividing (A1) by $dr$ and letting $x = (\phi\Delta\tau/i\, dr)^{-1}$, we have an expected marginal cost of benefits of $((1 - e^{-i})/i)^2(e^x - 1)/x$. Now $x \to 0$ as $dr \to 0$, and so as increments to unemployment become small this marginal cost approaches

$((1-e^{-i})/i)^2 \approx 1/(1+i)$. Therefore, the reserve-ratio system generates approximately a one-year interest-free loan for *UI* payments made by firms located in large steps of the tax function. Consequently, the marginal cost of benefits in reserve-ratio states jumps from $\eta_1/(\eta_1+i)$ to $1/(1+i)$ for equilibrium unemployment rates in the range $(\tau_0/\rho, \tau_1/\rho)$.

### B. *The Likelihood Function*

In a sample of full-time full-year participants, individuals may be categorized among three labor force "states" as of the sample period, $t$. These are employment ($e$), on temporary layoff ($l$), and unemployed without prospect of rehire ($d$). Over time, individuals move among these states, and I assume that in a stationary environment the cumulative distribution of *completed* spell lengths $\tau$ in state $i$ is $F^i(\tau)$. The conditional density of *incomplete* spell lengths among workers in state $i$ at the survey date is then

$$\text{(A2)} \quad h^i(T)=\frac{v^i(t-T)}{V^i(t)}\{1-F^i(T)\},$$

where $v^i(t-T)$ is the probability that a spell commenced at $t-T$ and $V^i(t)$ is the probability of observing an individual in $i$. The probability of observing the $T$th week of a spell of type $i$ is then simply $h^i(T)V^i(t)$.

Now if transitions are first-order Markov, a well known result from renewal theory (D. R. Cox, 1962) states that the distribution functions $F^i(\tau)$ must be exponential. (For a more general analysis, see Chris Flinn and James Heckman, 1980.) The conditional exit-time densities given elapsed time in the state—the hazard functions—are constants given by $f^i(\tau)/(1-F^i(\tau))$. Denote these transition rates by $\lambda^{el}$, $\lambda^{ed}$, $\lambda^{le}$, and $\lambda^{de}$. I assume that $\lambda^{dl}=\lambda^{ed}=0$, so that permanent layoffs cannot become temporary and conversely.

Straightforward calculation shows that the steady-state probability of observing randomly selected individual $a$ in state $e$, employment, is simply

$$\text{(A3)} \quad \phi^e(t)=\frac{\lambda^{le}}{\lambda^{el}+\lambda^{le}+\lambda^{ed}\lambda^{le}/\lambda^{de}},$$

which is independent of $t$. Using this, we have, for example, $v^l(t-\tau)=\lambda^{el}\phi^e(t)$. Therefore, the probabilities of observing the

TABLE A1—VARIABLE DEFINITIONS AND SUMMARY STATISTICS, MARCH 1975 *CPS* SAMPLE

| Variable | Definition | Mean | Standard Deviation |
|---|---|---|---|
| Subsidy | Total *UI* subsidy as proportion of weekly earnings (imputed) | .31 | .19 |
| Replacement Ratio | Potential *UI* benefits as proportion of weekly after-tax earnings (imputed) | .56 | .15 |
| Growth Rate[a] | Predicted rate of growth of state, private, non-agricultural employment from auxiliary regressions | 1.27 | 1.86 |
| Weekly Wage | Annual earnings ÷ weeks worked last year | 202.49 | 98.02 |
| Age | Age in years | 39.50 | 12.60 |
| Education | Years of completed schooling | 12.20 | 2.50 |
| Sex | = 1 if individual is male, = 0 otherwise | .73 | .45 |
| Race | = 1 if individual is white, = 0 otherwise | .90 | .31 |
| $T$ | Duration of a temporary layoff spell in progress (weeks) | 10.50 | 7.50 |
| Layoffs[a] | = 1 if worker on temporary layoff, = 0 otherwise | 7.11 | 25.70 |
| Observations | | 7,806 | |

*Note*: Original sample included 8,280 observations. The estimation procedure conditions on the sample of workers who are on layoff or employed, so discharges (435) and quits (39) are deleted from the final sample.

[a] Growth Rate and Layoffs are shown in percent.

$T$th week of a layoff and discharge spell are, respectively,

$$\text{(A4)} \quad \phi^l(t,T) = \phi^e(t)\lambda^{el}\exp\{-T\lambda^{el}\},$$

$$\text{(A5)} \quad \phi^d(t,T) = \phi^e(t)\lambda^{ed}\exp\{-T\lambda^{ed}\}.$$

These depend on all four hazards, and so estimation of the full model involves a number of parameters equal to four times the number of regressors. We may economize to analyze passages to and from temporary layoff, however, by conditioning the sample on those individuals who are not in state $d$. Using (A3) and (A4) this reduces the number of unknowns to $\lambda^{le}$ and $\lambda^{el}$, and results in equations (2) and (3) of the text. This sample restriction reduced the sample size to 7,806 from 8,280.

Summary statistics for these data used in estimating this model are shown in Table A1.

## REFERENCES

**Baily, Martin Neil,** "On the Theory of Layoffs and Unemployment," *Econometrica*, July 1976, *44*, 1043–63.

**Becker, Joseph,** *Experience Rating in Unemployment Insurance: An Experiment in Competitive Socialism*, Baltimore and London: Johns Hopkins University Press, 1972.

______, *Unemployment Insurance Financing*, Washington: American Enterprise Institute, 1981.

**Brechling, Frank,** "Layoffs and Unemployment Insurance," in S. Rosen, ed., *Studies in Labor Markets*, Chicago: University of Chicago Press, 1981.

**Cox, D. R.,** *Renewal Theory*, London: Methuen and Co., 1962.

**Diamond, Peter,** "Mobility Costs, Frictional Unemployment, and Efficiency," *Journal of Political Economy*, August 1981, *89*, 798–812.

**Feldstein, Martin,** "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers on Economic Activity*, 3: 1975, 725–44.

______, "Temporary Layoffs in the Theory of Unemployment," *Journal of Political Economy*, August 1976, *84*, 937–57.

______, "The Effect of Unemployment Insurance on Temporary Layoff Unemployment," *American Economic Review*, December 1978, *68*, 834–46.

______, "The Economics of the New Unemployment," *Public Interest*, Fall 1973, *33*, 3–42.

**Flinn, Chris and Heckman, James,** "Models for the Analysis of Labor Force Dynamics," Working Paper, National Opinion Research Center, Chicago, 1980.

**Hall, Robert,** "Turnover in the Labor Force," *Brookings Papers on Economic Activity*, 3: 1972, 709–56.

______, "A Theory of the Natural Unemployment Rate and the Duration of Unemployment," *Journal of Monetary Economics*, April 1979, *5*, 153–69.

**Hamermesh, Daniel,** *Jobless Pay and the Economy*, Baltimore: Johns Hopkins University Press, 1979.

**Kiefer, Nicholas and Neumann, George,** "An Empirical Job-Search Model, with a Test of the Constant Reservation-Wage Hypothesis," *Journal of Political Economy*, February 1979, *87*, 89–107.

**Nickell, Steven,** "Estimating the Probability of Leaving Unemployment," *Econometrica*, September 1979, *47*, 1249–66.

**Salant, Steven,** "Search Theory and Duration Data: A Theory of Sorts," *Quarterly Journal of Economics*, February 1977, *91*, 39–57.

**Topel, Robert,** (1982a) "Inventories, Layoffs, and the Short-Run Demand for Labor," *American Economic Review*, September 1982, *92*, 769–87.

______, (1982b) "Unemployment Insurance, Experience Rating, and the Occurrence of Unemployment," unpublished paper, University of Chicago, January 1982.

______ **and Welch, Finis,** "Unemployment Insurance: Survey and Extensions," *Economica*, August 1980, *47*, 351–79.

**Welch, Finis,** "What Have We Learned from Empirical Studies on Unemployment Insurance?," *Industrial and Labor Relations Review*, July 1977, *30*, 451–61.

**U.S. Department of Labor, Bureau of Labor Sta-**

tistics, *Employment and Earnings, States and Areas*, Washington: USGPO, 1909–75.

______, *Current Population Survey*, March File (computer tape), 1975.

______, *Unemployment Insurance Statistics*, Washington, 1979.

______, **Bureau of Employment Security**, *Significant Provisions of State Unemployment Insurance Laws*, Washington, January 1979.

______, **Employment and Training Administration**, *Handbook of Unemployment Insurance Financial Data [1938–76]*, Washington, 1978.

______, **Unemployment Insurance Service**, *Handbook of Unemployment Insurance Financial Data [1938–1980]*, Washington 1971.

# The Origin of Predictable Behavior

*By* RONALD A. HEINER*

Despite vigorous counterargument by its proponents, optimization theory has been persistently attacked as an acceptable explanation of behavior. In one form or another, these attacks repeat the oldest critique of economics; namely, the ability of agents to maximize successfully. Over the years, this critique has taken various forms which include information processing limitations in computing optima from known preference or utility information, unreliable probability information about complex environmental contingencies, and the absence of a well-defined set of alternatives or consequences, especially in an evolving world that may produce situations that never before existed.

These complaints are not new to economics. Indeed, they have been present during the very intellectual sifting process that produced neoclassical optimization and general equilibrium theory. Thus, if we are to further elaborate this critique of conventional theory, the basic issue is whether there is anything new that is worthy of attention by someone well versed in standard tools and concepts. Are we simply advancing more refined or cleverly argued versions of older critiques, or extensions of them to areas not previously emphasized?

Such arguments would still represent an attack on the basic rationality postulate of economics (that agents are able to maximize), but without providing a clear alternative to traditional optimization theory. However plausible these arguments might be, ultimately they must be set aside by someone desiring a theoretical understanding of behavior, unless they lead to another modeling structure whose analytical ability can be explored and compared with existing optimization theory.

Another argument focuses on the desire to understand the "real" dynamic processes that actually generate observed behavior. In contrast, optimization is thought of as a surrogate theory based on false assumptions about agents' capacity to maximize. Thus, it can be defended only in terms of empirical testability, without really illuminating the underlying processes determining behavior.

Nevertheless, even if this view was fully accepted, it is unlikely by itself to cause a major shift away from conventional thinking. The reason is that evolutionary processes have long ago been interpreted as one of the key mechanisms tending to produce optimizing behavior; or conversely, optimizing models will predict the behavior patterns that will survive in an evolutionary process tending to select relatively superior performance.[1] The latter interpretation is in fact one of the dominant justifications for standard models against the criticism of unrealistic assumptions (i.e., the surviving agents of a selection process will behave "as if" they are able to maximize).[2]

---

*Department of Economics, Brigham Young University, Provo, UT 84602. I am indebted to Axel Leijonhufvud for constant encouragement about applications to economics, and for numerous stylistic suggestions. Harold Miller helped familiarize me with a broad range of issues across the sociobiological, psychological, and behavioral science literatures. James Buchanan provided stimulating discussion about conceptual issues. I have also benefited from the advice and criticism of Armen Alchian, Ron Batchelder, Bruce Brown, Robert Clower, Daniel Friedman, Jack Hirshleifer, Kai Jeanski, Randy Johnson, Edward Leamer, Stephen Littlechild, John McCall, James McDonald, Richard Nelson, Gerald O'Driscoll, Dennis Packard, Clayne Pope, Lionello Punzo, Ezio Tarantelli, and Sidney Winter. Needless to say, these colleagues are not responsible for inadequacy in the conceptual framework or scope of ideas presented.

[1] See in particular Armen Alchian's well-known 1950 paper, and also Sidney Winter, 1964, 1971; Jack Hirshleifer, 1977; Richard Nelson and Winter, 1974.

[2] A still used reference on the "as if" point of view is Milton Friedman's 1953 paper. Some recent journal illustrations are Benjamin Klein and Keith Leffler, 1981, p. 634; Richard Posner, 1980, p. 5; Hirshleifer, 1977, p. 50; Nelson, 1981, p. 1059. The ultimate extension of this view is to claim not that agents are able to maximize (select most preferred actions), but rather that any ob-

In spite of the above conclusions, I believe there is a viable alternative to standard models—one that directly comes to grips with the persistent critiques of economic theory and which broadens our analytical horizon to encompass a much wider range of phenomena.

In particular, I believe that observed regularities of behavior can be fruitfully understood as "behavioral rules" that arise because of uncertainty in distinguishing preferred from less-preferred behavior. Such uncertainty requires behavior to be governed by mechanisms that restrict the flexibility to choose potential actions, or which produce a selective alertness to information that might prompt particular actions to be chosen. These mechanisms simplify behavior to less-complex patterns, which are easier for an observer to recognize and predict. In the special case of no uncertainty, the behavior of perfectly informed, fully optimizing agents responding with complete flexibility to every perturbation in their environment would not produce easily recognizable patterns, but rather would be extremely difficult to predict. Thus, it is in the limits to maximizing that we will find the origin of predictable behavior.

If the view taken here is correct, it means that predictable features of behavior do not arise from optimizing with no uncertainty in choosing most preferred behavior; and furthermore, evolutionary selection processes will in general not produce approximations to optimizing behavior. Rather, predictable behavior will evolve only to the extent that uncertainty prevents agents from successfully maximizing.

In the following, I sketch the line of thought and the observations which have led me to this conclusion, and briefly outline the elements of a modeling structure that can be applied to a wide range of topics. A number of applications are presented to illustrate the range of issues unified by the analysis, which is far broader than the reader is likely to anticipate without explicit examples.

## I. Problems with the Methodological Arguments for Optimization

Optimizing with full ability to select most preferred behavior is rarely justified as an empirically realistic assumption. Rather, it is usually defended on methodological grounds as the appropriate theoretical framework for analyzing behavior. The chief defense is empirical fruitfulness in generating unfalsified predictions.

We might criticize this testability criteria with modern philosophy of science arguments.[3] Nevertheless, a long list of confirmed predictions would be persuasive evidence in favor of a theory. Yet, it is just here that we have a problem. Suppose we really asked to see the list of clearly implied, unambiguous predictions that have been derived from our basic optimization models.

The answer to this query, one that would be admitted by many practitioners in the field, is that at best we have developed a very short list. All sorts of behavior is consistent with or plausibly suggested by optimization models, yet still not predicted by them. For example, optimization models have never been able to imply the Law of Demand (buying less of a commodity when its price rises), which is probably the oldest and simplest behavioral regularity in economics. Of course, we can use the theory to argue it is unlikely that a negative income effect will outweigh the pure substitution effect, especially for goods that absorb a small fraction of a person's income.[4] The acceptance of this view is heavily influenced by our belief in the

---

served behavior is consistent with the maximization of some function. This latter formulation is probably incapable of either theoretical or empirical disproof (see Lawrence Boland, 1981).

[3] See for example, B. Caldwell, 1982; also Karl Popper, 1969; Imer Lakatos and Alice Musgrave, 1970.

[4] I was told in a graduate price-theory class by Armen Alchian that the only clear implication of consumer theory is that with more income, a consumer will buy more of at least something. Harold Demsetz, when informed of this story, responded by saying, "well then just define holding cash balances as saving, and we have no testable implications, just one mass of tautologies." See also the opening remarks of Kenneth Arrow, 1982, p. 1; and the closing remarks of Vernon Smith, 1982, p. 952.

empirical validity of the Law of Demand. Yet, regardless of how cleverly we interpret a Slutsky equation, no clear prediction is implied.

We could pursue a number of other examples, all of which suggest that conventional models have never really been fruitful in generating testable implications.[5] For this reason, I believe allegiance to these models is not grounded in the claim of empirical fruitfulness, despite the usual rhetoric that this is the case. Rather, it is based on a deeper methodological issue about the effect of dropping the basic rationality assumption.[6]

Think of this issue in the following terms. Standard choice theory tries to explain behavior by matching the "competence" of an agent with the "difficulty" in selecting most preferred alternatives. It assumes for the purpose of theoretical explanation that there is no gap between an agent's competence and the difficulty of the decision problem to be solved (hereafter called a "*C-D* gap").[7] On the other hand, the presence of a *C-D* gap will introduce uncertainty in selecting most preferred alternatives, which will tend to produce errors and surprises. Such mistakes are by their nature unpredictable and erratic. Yet, it is only the systematic elements of behavior that we can hope to scientifically explain and predict. Thus, in order to theoretically isolate the systematic tendencies in behavior, we must exclude a *C-D* gap, no matter how implausible or unrealistic this might be.[8]

This perspective has been a dominant factor in loyalty to traditional optimizing concepts. Nevertheless, I believe it is mistaken, and that essentially the opposite view is true. To see why, think of the above argument as an empirical hypothesis about the effect of "irrationality"; namely, that the additional uncertainty from a larger *C-D* gap will generate more errors and surprises, thus producing more irregularity and noise in behavior. There are numerous complicating factors about how to test this hypothesis, especially how to measure a person's *C-D* gap. We can avoid these problems by broadening our horizon to consider an interspecie comparison between humans and other animals. Here it is clear without detailed argument that the average *C-D* gap of other animals is larger than that of humans.[9] Yet when we observe nonhuman species, the overwhelming qualitative impression is not one of greater irregularity, but instead of greater rigidity and inflexibility of behavior. Pattern is not more difficult but rather easier to notice in animals than in humans.

---

[5] Some other examples are: second Law of Demand, short- and long-run supply dynamics, risk aversion, time preference, self-interest, liquidity preference, expectation lag and adjustment structures, price-taking behavior, oligopoly strategic patterns, relative price vs. quantity elasticities, relative income-consumption elasticities, etc. The so-called "laws of supply and demand" have probably been the most empirically useful tools in economics (both in formulating simple hypotheses about market responses to parameter changes, and in providing the basic structural equation system used in modern econometric model building). Yet, these simple laws are not derivable from basic optimization concepts, and thus empirical analysis derived from them does not confirm these concepts.

[6] Without going into any details, I would also like to mention a large literature in behavioral psychology about the *matching law* (Richard Herrnstein, 1961, 1964, 1970), which has cast doubt on the validity of traditional maximization theory to explain behavior under certain reinforcement schedules. See P. de Villiers, 1977, for a summary of earlier experimental results, and for more recent experiments with human subjects, see C. M. Bradshaw, E. Szabadi, and R. Bevan, 1976; William Buskist and Harold Miller, 1981. For recent articles about the validity of maximization, see Herrnstein and Gene Heyman, 1979; Heyman and R. Duncan Luce, 1979; Howard Rachlin, John Kagel and R. C. Battalio, 1980; D. Prelec, 1982; and for recent experiments in which matching has dominated maximizing behavior, see Herrnstein and William Vaughan, 1980; Vaughan, 1981; John Mazur, 1981.

[7] Posing the problem in terms of a gap in an agent's decision competence relative to the difficulty of a decision problem was suggested to me by Axel Leijonhufvud.

[8] For a recent example of this view, see Jack Hirshleifer's 1980 price theory text, p. 9. A similar argument is used to justify "rational expectations" equilibria. See, for example, Robert Lucas, 1981, pp. 125, 223–24; and Robert Cooter, 1982b, p. 232.

[9] For analysis of cognative differences between humans and animals, and the evolution of intelligence, see M. Konner 1982, David Premack, 1983; P. Rozin, 1976; Carl Sagan, 1977; Harry Jerison, 1973; R. Masterton, William Hodos, and Jerison, 1976.

This qualitative difference between humans and other animals is obviously not new to us; it having long ago been given the capsulized description of "instinct." Still, I do not believe that we have recognized the significance of this general pattern for evaluating and constructing theoretical models of behavior. This pattern is telling us that it is not the absence of a *C-D* gap, but rather its presence which conditions regularity in behavior.

Why should this be the case? Think of an omiscient agent with literally no uncertainty in identifying the most preferred action under any conceivable condition, regardless of the complexity of the environment which he encounters. Intuitively, such an agent would benefit from maximum flexibility to use all potential information or to adjust to all environmental conditions, no matter how rare or subtle those conditions might be. But what if there is uncertainty because agents are unable to decipher all of the complexity of the environment (i.e., there is uncertainty due to a *C-D* gap)? Will allowing complete flexibility still benefit the agents? For example, if we could somehow "loosen up" the behavior of an organism without affecting its perceptual abilities, would it compete more effectively for food or mating partners than before?

I believe the general answer to this question is negative: that when genuine uncertainty exists, allowing greater flexibility to react to more information or administer a more complex repertoire of actions will not necessarily enhance an agent's performance. Even if we confine our attention to human behavior, we can find evidence for this proposition, especially in highly competitive situations with noticable elements of complexity relative to human information processing and other perceptual abilities.

For example, in sequential replication games of the basic prisoner's dilemma (see Robert Axelrod, 1980a), round robin competition identified the simplest strategy (the tit for tat strategy) as dominant over all of the others (submitted by persons in economics, mathematics, psychology, political science, and sociology).[10] Moreover, the worst performance came from the strategy that specified the most "sophisticated" learning and probability adjustment process to guide its behavior.[11] Another example is the publishing history on strategies to win at blackjack. Earlier books emphasized sophisticated card-counting, bet-variation methods (see especially Edward Thorpe's book, *Beat the Dealer*). However, while no one has challenged the mathematical validity of these earlier more complex methods, their actual use resulted in worse performance by most persons attempting to use them (which generated sizable unexpected profits to the casinos).[12] As a result, later books have steadily evolved toward more rigidly structured methods (for example, two recent books are *No Need to Count* and *Winning Casino Blackjack for the Non-Counter*).[13]

Consider also Rubic's cube. There are over 43 trillion possible initial positions from which to unscramble the cube. Minimizing the number of moves to solve the cube would require an extremely complex pattern of adjustment from one particular scrambled position to another. Yet, if mistakes are made in trying to select a short cut, the cube will remain unscrambled indefinitely. Consequently, cube experts have developed rigidly structured solving procedures that employ a small repertoire of solving patterns to unscramble the cube. These procedures follow a predetermined hierarchical sequence that is largely independent of the initial scrambled position.[14] However, they almost always require a much longer sequence of moves than the minimum number needed to unscramble the cube. Thus, they are not an approximation to the enormously complex behavior that would be exhibited by an omniscient agent who could immediately select the shortest sequence for each scrambled position. Note also that the information needed to behave in this fashion (present in the initially scrambled patterns on the face of the cube) is costless to observe and instantly available; one need only look at the cube while unscrambling it.

---

[10] For a description of the tournament and its results, see Axelrod, 1980a. The top strategies were all variants of the simple tit for tat strategy, but none were able to beat the basic strategy (in particular, see pp. 8 and 18). When a second round of the tournament was run, tit for tat still won even though numerous more complex strategies were submitted (see Axelrod, 1980b). For recent analytical analysis on this issue, see David Kreps et al., 1982.

[11] Axelrod describes the worst of the submitted strategies in the first round:

> This rule has a probability of cooperating, *P*, which is initially 30% and is updated every 10 moves. *P* is adjusted if the other player seems random, very cooperative, or very uncooperative. *P* is also adjusted after move 130 if the rule has a lower score than the other player. Unfortunately, the complex process of adjustment frequently left the probability of cooperation in the 30% to 70% range, and therefore the rule appeared random to many other players. [1980a, p. 24]

[12] For example, see Richard Canfield, 1979, pp. 19, 37–38, 144–47, 150.

[13] Some of the major books in order of publication are: Thorpe, 1962; Lawrence Revere, 1969; John Archer, 1973; Ian Anderson, 1975; Virginia Graham and C. I. Tulcea, 1978; Canfield, 1979; Leon Dubey, 1980; Avery Cardoza, 1981. See Canfield's book, especially pp. 11–12, 16–19, 37–38, 60–61, 62–65. See also Dubey, pp. 11–12, 17–19, 64, 165–66, 168, 172.

Finally, consider the research of Herbert Simon over a number of years,[15] which has shown that decision makers in a variety of contexts (including both individual and organizational behavior) systematically restrict the use and acquisition of information compared to that potentially available. For example, Simon's idea of "satisficing" represents a feedback mechanism between an internal target variable (called the "aspiration level") and the scope of information evaluated to implement that target. Over time, the feedback process will both guide and discipline the use of information and the resulting behavioral complexity that will evolve within a person or organization. Other learning, cognitive processes, and decision algorithms can be similarly interpreted.

The above examples suggest that allowing flexibility to react to information or to select actions will not necessarily improve performance if there is uncertainty about how to use that information or about when to select particular actions. Thus, an agent's overall performance may actually be improved by restricting flexibility to use information or to choose particular actions.

## II. How Uncertainty Generates Flexibility Constrained Behavior

The argument to this point has suggested that uncertainty due to a *C-D* gap may generate flexibility constrained behavior. The next step is to characterize more precisely how such uncertainty might produce this result. To do so, a simple "reliability condition" is developed that specifies when to allow or prohibit flexibility to select potential actions or to use information that might prompt particular actions to be chosen.

Two major classes of variables determine the uncertainty resulting from a *C-D* gap. The first are environmental variables (denoted by **e**) which determine the complexity of the decision problem to be solved by an agent (including the complexity of environmental situations potentially encountered; the relative likelihood of these situations; and the stability of the relationships that determine possible situations and their relative likelihood). The second are perceptual variables (denoted by **p**) which characterize an agent's competence in deciphering relationships between its behavior and the environment.[16] Thus, the **p** and **e** variables determine the "gap" between competence and difficulty (the *C-D* gap) which produces uncertainty about how to use information in selecting potential actions. In general, there is greater uncertainty as either an agent's perceptual abilities become less reliable or the environment becomes more complex.

---

[14] In following a typical set of instructions, one selects a side of the cube and begins by placing either its corner or its edge pieces in their proper positions; next, one places in sequence the pieces in the middle section; finally, one repositions the pieces on the remaining, opposite side of the cube (see Czes Kosniowski, 1981). Other similar procedures include D. Taylor, 1980; James Nourse, 1980; Patrick Bussert, 1981; B. W. Barlow, 1981.

[15] For example, Simon, 1955, 1959, 1969, 1976, 1978, 1979a; A. Newell and Simon, 1972.

[16] In economics, the **p** variables might describe mistaken perceptions about what is more preferred, information processing errors, unreliable probability information, etc.; while the **e** variables describe the complexity and volatility of both present and future exchange, legal, and political conditions. In biology, **p** might refer to the sensory and cognitive mechanisms of an organism, and **e** to the structure and stability of ecological relationships involving competition for food or mating partners.

These relationships are formally represented as a vector-valued function, $\mathbf{U} = u(\overset{-}{\mathbf{p}}, \overset{+}{\mathbf{e}})$, which describes the structure of uncertainty from a *C-D* gap characterized by **p** and **e**. The signs above **p** and **e** signify that uncertainty is negatively related to an agent's perceptual abilities, and positively related to the complexity and instability of the environment.

Now consider a conceptual experiment about an agent initially limited to a fixed repertoire of actions, and ask whether allowing flexibility to select an additional action will improve the agent's performance. Under certain conditions, the new action will be more preferred than the other actions in the agent's repertoire (the "right" time to select the action), but otherwise it will be less preferred than one of those actions (the "wrong" time to select the action). Depending on the likelihood of different situations produced by the environment, the probabilities of the right or wrong time to select the action are written as $\pi(\mathbf{e})$ and $1-\pi(\mathbf{e})$, respectively.

Because of uncertainty, the agent will not necessarily select the new action when it is the right time to do so. The conditional probability of selecting the action when it is actually the right time is written $r(\mathbf{U})$, where the likelihood of so doing depends on the structure of uncertainty, $\mathbf{U} = u(\mathbf{p}, \mathbf{e})$. When this happens, the resulting gain in performance (compared to staying within the initial repertoire) is written $g(\mathbf{e})$, which depends on how the environment affects the consequences from different actions. Similarly, the conditional probability of selecting the new action when it is actually the wrong time is written $w(\mathbf{U})$, with consequent loss in performance of $l(\mathbf{e})$.

In the special case of no uncertainty, the new action would always be selected at the right time and never at the wrong time, so that $r = 1$ and $w = 0$. In general, however, the presence of uncertainty will imply $r < 1$ and $w > 0$.

We can intuitively measure the *reliability* of selecting a new action by the ratio $r/w$, which represents the chance of "correctly" selecting the action at the right time relative to the chance of "mistakenly" selecting it at the wrong time.[17] Greater uncertainty will both reduce the chance of correct selections and increase the chance of mistaken selections, thus causing the ratio $r/w$ to drop (i.e., greater uncertainty reduces the reliability of selecting the new action).

Note also that $r(\mathbf{U})$ and $w(\mathbf{U})$ are not assumed to be known to an agent. The reason is that uncertainty produces mistakes about distinguishing the right from the wrong conditions to select an action, which distinction is necessary to determine the conditional probabilities of choosing an action under these two sets of conditions. For the same reason, the probability of the right situation to select an action, $\pi(\mathbf{e})$, may also be unknown to an agent. Thus, it is not assumed that an agent can tell whether a mistake has been made; nor are we necessarily dealing with situations where an agent consciously decides when to select an action. Rather, the more general issue is whether some process —conscious or not—will cause (or prevent) an "alertness" or "sensitivity" to information that might prompt selection of an action. For example, when will a person develop an alertness to potential information about whether to choose a particular action, or whether to modify a previous behavior pattern; or when will instinctive mechanisms in an organism precondition a sensitivity to certain environmental stimuli, while simultaneously blocking alertness to other potential stimuli.

---

[17] The probabilities $r$ and $w$ can also be interpreted using Type 1 and Type 2 errors used in statistical hypothesis testing. Let the null hypothesis represent the right situation to select an action (when it is more preferred); while the alternate hypothesis represents the wrong situation for selecting it. Thus, intuitively, Type 1 errors represent *excluded benefits* from failing to respond under the right conditions, while Type 2 errors refer to *included mistakes* from still responding under the wrong conditions. If we let $t_1$ and $t_2$ denote the respective probabilities of these errors, they characterize $r$ and $w$ by $r = 1 - t_1$, and $w = t_2$. Thus, $r$ equals one minus the chance excluded benefits, and $w$ equals the change in included mistakes.

Now, with the above components, we can formulate an answer to the question posed earlier: *when is the selection of a new action sufficiently reliable for an agent to benefit from allowing flexibility to select that action.*

To answer this question we must determine whether the gains $g(\mathbf{e})$ from selecting the action under the right conditions (when it is actually more preferred) will cumulate faster than the losses $l(\mathbf{e})$ from selecting it under the wrong conditions (when it is actually less preferred). Thus, combine the above elements in the following way. Right conditions occur with probability $\pi(\mathbf{e})$, which are correctly recognized with probability $r(\mathbf{U})$; so that the expected gain from allowing flexibility to select another action is $g(\mathbf{e})r(\mathbf{U})\pi(\mathbf{e})$. Similarly, the expected loss conditional on allowing the action to be selected is $l(\mathbf{e})w(\mathbf{U})(1-\pi(\mathbf{e}))$. Accordingly, gains will cumulate faster than losses if $g(\mathbf{e})r(\mathbf{U})\pi(\mathbf{e}) > l(\mathbf{e})w(\mathbf{U})(1-\pi(\mathbf{e}))$. Hence, simple rearrangement yields the following *Reliability Condition*:

$$\frac{r(\mathbf{U})}{w(\mathbf{U})} > \frac{l(\mathbf{e})}{g(\mathbf{e})} \cdot \frac{1-\pi(\mathbf{e})}{\pi(\mathbf{e})}.$$

The left-hand side of the inequality is a *reliability ratio*, $r(\mathbf{U})/w(\mathbf{U})$, which measures the probability of "correctly" responding under the right circumstances relative to the probability of "mistakenly" responding under the wrong circumstances. The right-hand side of the inequality represents a minimum lower bound or *tolerance limit* (hereafter denoted simply by $T(\mathbf{e}) = l(\mathbf{e})/g(\mathbf{e}) \times (1-\pi(\mathbf{e}))/\pi(\mathbf{e}))$, which a reliability ratio must satisfy. That is, $T(\mathbf{e})$ determines how likely the chance of selecting an action under the right conditions must be compared to the chance of selecting it under the wrong conditions before allowing flexibility to select that action will improve performance.

We can intuitively interpret the ratio $r(\mathbf{U})/w(\mathbf{U})$ as the "actual" reliability of selecting an action, in comparison to the minimum "required" reliability specified by the tolerance limit, $T(\mathbf{e})$. The components of the Reliability Condition summarize a potentially complex set of relationships between an agent's repertoire and the structure of the environment.[18] Nevertheless, these relationships boil down to a conceptually simple answer about when to allow flexibility to select an additional action: *do so if the actual reliability in selecting the action exceeds the minimum required reliability necessary to improve performance.* Stated in its simplest notational form, this answer amounts to the condition, $r/w > T$.

The question which motivated this answer was phrased in terms of adding a new action to an agent's repertoire. However, once the Reliability Condition has been obtained we can also apply it to a range of further issues about when to allow or ignore particular actions. For example, it can be applied to dropping actions from a repertoire; namely, retain only those actions which satisfy $r/w > T$ compared to ignoring them.

We can also think of the Reliability Condition as solving a "decision" problem in which an agent determines what information he will allow to influence his behavior; or alternatively, as a "design" problem in engineering the appropriate information sensitivity of an agent. For each possible action, the Reliability Condition must be satisfied before allowing potential information to prompt its selection. Those actions that can be guided with sufficient reliability are permitted; those that cannot are eliminated. In this way, an agent's outward behavior is determined by his response pattern to potential information.[19]

---

[18] Both the agent's repertoire and the environment may contain a large number of possibilities, and the consequences from selecting an action may vary with different environmental situations. This will also complicate how to measure an agent's performance. Regardless of how performance is measured (for example, it may involve some kind of average over actions and/or environmental conditions), $g(\mathbf{e})$ and $l(\mathbf{e})$ still represent the gain or loss in performance from correct or mistaken selections, respectively; and $r(\mathbf{U})$, $w(\mathbf{U})$ still represent the conditional probabilities of these correct or mistaken selections. The probabilities $r(\mathbf{U})$ and $w(\mathbf{U})$ also result from a complex set of relationships that determine the source and likelihood of particular errors that interact to generate these probabilities. In addition, $l(e)$ and $g(e)$ may depend on an agent's internal components, such as the morphological attributes of an animal.

The objective of this paper is to develop only the bare essential modeling elements needed for a simple analytical solution, whose structure is invariant to the above-mentioned complications. In particular, the basic form of the Reliability Condition will remain the same. Much greater detail about the analytical structure, including extensive applications to economics and other fields, is now in progress.

## III. Four General Implications

Now that we have the Reliability Condition, its implications in four basic areas are briefly discussed.

### A. *Uncertainty Generates Rules Which are Adapted Only to Likely or Recurrent Situations*

Note a simple but important feature of the tolerance limit. For any given $l/g$ ratio, the likelihood of wrong to right conditions, $(1-\pi)/\pi$, increases for smaller $\pi$; so that $T$ also rises as the probability of right circumstances $\pi$ decreases (see Figure 1). Thus, *an agent must be more reliable in selecting an action if the right situations for exhibiting it are less likely. Moreover, the required reliability quickly accelerates to infinity as the likelihood of right situations drops to zero.* Thus, for a given structure of uncertainty, $\mathrm{U}=u(\mathbf{p},\mathbf{e})$, which determines the reliability of selecting a particular action (i.e., which determines the ratio $r(\mathrm{U})/w(\mathrm{U})$), the Reliability Condition will be violated for sufficiently small but positive, $\pi(\mathbf{e})>0$.

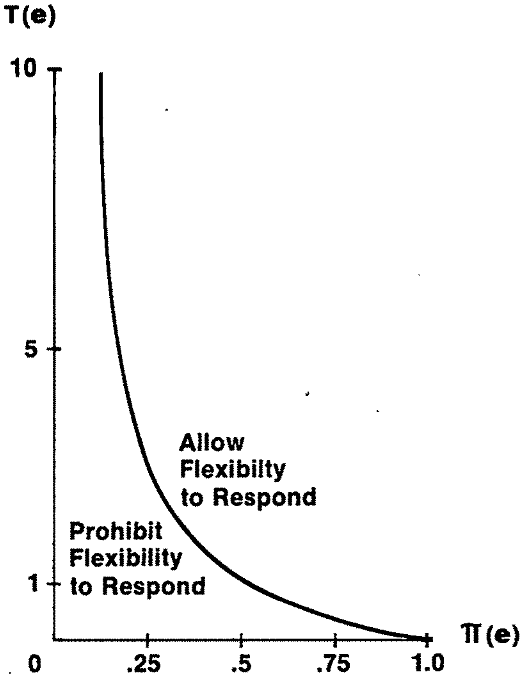


FIGURE 1

The curve shows how the tolerance limit $T(e)$ changes for a constant $l(e)/g(e)$ ratio (in this case $l/g=1$) as the probability of right conditions $\pi$ varies. Note how quickly $T$ begins to rise as $\pi$ drops below .25. The curve represents a boundary of minimum reliability that must be satisfied (i.e., $r/w>T$) before responding to information will enhance an agent's performance.

This intuitively means that to satisfy the Reliability Condition, an agent must ignore actions which are appropriate for only "rare" or "unusual" situations. Conversely, an agent's repertoire must be limited to actions which are adapted only to relatively likely or "recurrent" situations. Thus, a general characteristic of such a repertoire is that it excludes actions which will in fact enhance performance under certain conditions, even though those conditions occur with positive probability, $\pi(\mathbf{e})>0$. We thus have a formal characterization of the pervasive association of both human and animal behavior with various connotations of "rule-governed" behavior, such as instinct, habits, routines, rules of thumb, administrative procedures, customs, norms, and so forth. All of these phrases refer to some type of rigidity or inflexibility in adjusting to different situations as a universal qualitative feature of behavior.

Therefore, since behavior patterns which satisfy the Reliability Condition must have this property, we will call them *behavioral rules* or simply *rules*. Note that we have been able to derive the basic rigidity feature which justifies attributing to such behavior patterns the idea of rules. This contrasts sharply with the typical procedure of using the language of rules (often with the intent of suggesting certain connotations to the reader), yet without really justifying from a more basic theoretical structure why such terminology is appropriate.

---

[19] The relationship between information sensitivity and output complexity is also recognized in cybernetics; see Norbert Weiner, 1948, and W. Ashby, 1956. A reference in organizational behavior that refers to this is Barry Staw, Lana Sanderlands and Jane Dutton: "... a fundamental principle of cybernetics..., the number of output discriminations of a system (i.e., its behavioral repertoire) is limited by the variety of information inherent in its input" (1981, p. 517).

If we use the jargon of standard economics, rule-governed behavior means that an agent must ignore actions which are actually preferred under certain conditions. Thus, as intuitively suggested above, the resulting behavior patterns are *not* an approximation to maximizing so as to always choose most preferred alternatives (i.e., behaving "as if" an agent could successfully maximize with no *C-D* gap).

In general, rules restrict behavior to only a limited repertoire of actions. Such restrictions do not assume an awareness of all the potential actions or information which are thereby implicitly ignored. Thus, no explicit decision about what potential actions to ignore is necessarily involved.

An agent need only be capable of determining when to select particular actions from a limited range of allowable alternatives. To do so does not require an ability to understand why the resulting behavior patterns evolved. This is obviously the case for animals, where we do not expect them to have an "intellectual awareness" of why they are programmed to exhibit certain behavior patterns. Yet even for humans, the general characteristic will be an inability to articulate a full understanding of why particular behavior patterns have arisen. This is implied even though human behavior is much more flexible than that of other species, and even though conscious mental processes are involved in most human behavior patterns.[20]

As a simple example involving human behavior, consider the solving methods for Rubic's cube mentioned above in Section I. The environment represents all of the different scrambled positions or "situations" which might eventuate on the face of the cube, of which there are over 43 trillion. If each situation is produced by a simple random draw from the set of possible situations, the probability $\pi$ of the right situation (the appropriate scrambled position) arising for any particular solving sequence is extremely low. Assuming the $l/g$ ratio (resulting from unscrambling the cube in greater or lesser time, or number of moves) is not close to zero, the required reliability for selecting each of these sequences will also be very high. Without this ability, the repertoire of solving patterns must be severely restricted in order to satisfy the Reliability Condition, and structured so that their use is largely independent of particular scrambled positions (i.e., they are adapted only to the recurrent features of the environment).

### B. *Selection Processes do not Simulate Optimizing Behavior*

Up to this point we have thought of performance simply in terms of an agent's "preferences" about the consequences of particular actions. Now generalize its meaning to represent any factor that determines whether behavior will continue or persist in the environment encountered by an agent. This might involve a preference evaluation, competition for profits or investment capital, or possibly biological determinants of physical survival or reproductive probability. Whatever the interpretation, we can apply the Reliability Condition to determine when allowing flexibility to use potential information or to select actions will improve rather than worsen performance.

Now suppose the actual process generating behavior is an evolutionary process that tends to select relatively superior performance at any point of time. From what has already been derived, this implies that such selection processes will tend to produce rule-governed behavior that is not an approximation to always selecting actions that maximize performance. Thus, in general, evolutionary processes will *not* generate simulations to optimizing behavior. Rather, they will tend to produce rules that systematically restrict the flexibility of behavior compared to that which would be exhibited by a full optimizer in the absence of uncertainty.

---

[20] For related comments about the legitimacy of standard psychotherapy practices, see Donald Campbell's 1975 presidential address to the American Psychological Association.

As mentioned earlier, this implication directly contradicts one of the dominant justifications for assuming agents are able to optimize. Predictable behavior is not an "as if" simulation to optimizing, but rather will evolve only to the extent that agent's are unable to maximize because of uncertainty.

Generalizing the meaning of performance also implies that we are not necessarily dealing with traditional economic agents, such as consumer, firm, worker, investor, etc. Rather, we can think of an agent as any system of interacting components. For example, a system might refer to biological entities such as individual organisms, species, ecological systems, or possibly to subsystems within organisms studied in physiology or molecular biology. Still other examples might be computers or other artificial cybernetic mechanisms.

Whatever the interpretation, the Reliability Condition characterizes when to allow flexibility to use information or select actions applicable to that interpretation. For example, we might apply it to the following situations: when is it the right time to unscramble Rubic's cube by starting from a middle section rather than from one of its outer sections; when is it the right time to purchase more of a particular commodity rather than other commodities; when is it the right time to search for additional price or quality information about potential future purchasing decisions; when is it the right time for an animal to deviate from its usual foraging strategies for food; when is it the right time to cooperate by helping other individuals (i.e., when is it the right time to be "altruistic" rather than "selfish");[21] when is it the right time to modify genetic information to perpetuate traits acquired in the lifetime of a particular organism (i.e., when is it the right time to use "Lamarkian" genetic transmission);[22] or more generally, when is it the right time to use feedback from the environment to modify behavior (i.e., when is it the right time to "learn")?

### C. *Weak Selection Processes May Allow Dysfunctional Behavior to Persist*

The preceding discussion implicitly assumed that selection processes would quickly eliminate relatively inferior performers. If this is actually the case, the Reliability Condition implies the evolution of behavioral rules that appropriately structure and limit the flexibility of behavior. The empirical examples that helped motivate the formal analysis also involved behavior produced in highly competitive conditions (i.e., biological competition for survival between nonhuman agents; strategies to win at blackjack, or in prisoner's dilemma games, or in Rubic's cube contests; organizations competing in exchange environments for profits or investment capital, etc.).

On the other hand, what if there is something about the environment that only sluggishly weeds out worse performers, or which only infrequently produces situations that severely punish vulnerable behavior. This possibility is fundamentally important when genuine uncertainty exists, because there is no magical element (empirically or in theory) to guarantee that only appropriately structured behavior will evolve. Indeed, the core assumption is literally the absence of ability to decifer all of the complexity of the environment; especially one whose very structure itself evolves over time.

Thus, consider an evolving world produced through a mixture of selective processes. These processes will have varying degrees of severity in reacting to differential performance between competing agents. Such a world will be a continual mixture of appropriately and inappropriately structured behavior. In some cases, weak selection processes may allow relatively dysfunctional behavior to persist: possibly with worse average performance than other agents; or with slowly dwindling performance over time; or with vulnerable performance that awaits only the next infrequent but severe test to challenge its further persistence in the environment.

---

[21] Hirshleifer uses "recognition coefficients" (which represent particular examples of the $r(U)$ and $w(U)$ probabilities) to determine the reliability of helping strategies in identifying other agents with altruistic traits (1982, pp. 26–29). See also W. D. Hamilton, 1964; John Maynard Smith, 1964; Robert Trivers, 1971.

[22] On the "irreversibility" of genetic translation, see Jacques Monod, 1972, pp. 104–17.

This is clearly a different view from trying to comprehend the world as continually tending toward optimizing behavior. Indeed, we may be able to explain major features about the structure, occurrence, and error patterns of dysfunctional behavior. Only one class of possibilities is mentioned here, and briefly reconsidered at the conclusion of this paper. In particular, we can analyze the pattern of vulnerable behavior arising from political institutions, especially in the form of dysfunctional complexity in trying to manipulate the outcomes resulting from exchange competition. Specific instances of this issue have had a long history in economics about the scope of government regulation, and the debate over discretionary vs. rigid monetary policy.[23]

### D. *Greater Uncertainty will Cause Rule-Governed Behavior to be More Predictable*

What is the effect of *greater* uncertainty on rule-governed behavior? In general, greater uncertainty (from either less reliable perceptual abilities or a more unpredictable environment) will both reduce the chance of recognizing the right situation to select an action, and increase the chance of not recognizing the wrong situation for selecting it. That is, greater uncertainty will both reduce $r(\mathrm{U})$ and increase $w(\mathrm{U})$, so that the reliability ratios, $r(\mathrm{U})/w(\mathrm{U})$, of particular actions will drop.

As these ratios drop, some of them may no longer exceed their respective tolerance limits, resulting in violations of the Reliability Condition. More violations will occur as uncertainty becomes more pervasive. Thus, greater uncertainty will cause behavioral rules to be more restrictive in eliminating particular actions or response patterns to potential information. This will further constrain behavior to simpler, less sophisticated patterns which are easier for an observer to recognize and predict. Therefore, *greater uncertainty will cause rule-governed behavior to exhibit increasingly predictable regularities, so that uncertainty becomes the basic source of predictable behavior.*

This is the most important implication of my analysis, one that has far-reaching implications across a diverse range of fields. It also has important implications for how we have been trying to model behavior. It implies that genuine uncertainty, far from being unanalyzable or irrelevant to understanding behavior, is the very source of the empirical regularities that we have sought to explain by excluding such uncertainty.[24] This means that the conceptual basis for most of our existing models is seriously flawed.

A major symptom of this has been the dominant tendency to model more complex decision problems by implicitly upgrading the competence of the agent to handle that complexity (so that traditional optimizing concepts can be used). For example, the number of decision alternatives or competing agents is increased, or complex probabilistic contingencies are introduced, or repercussions from future events are permitted, etc. Over the years this has resulted in the characterization of increasingly sophisticated,

---

[23] Another area involves differences in productivity between U.S. and Japanese industrial firms, because of differential ability either to manage a complex internal use of inputs, or to adjust to volatile external marketing conditions ("just in time" rather than "just in case" inventory management; greater employee discretion in production line monitoring; longer promotion, investment, and *R&D* planning horizons; etc.) See William Abernathy, Kim Clark, and A. Kantrow, 1981; Y. Mondon, 1981; Y. Sugimori, K. Kusunoki, and S. Cho, 1977; R. Clark, 1979; Anthony Athos and Richard Pascale, 1981; William Ouchi, 1981.

[24] The various authors that have emphasized the importance of uncertainty (for example, Frank Knight, 1921; the Australian view typified by F. A. Hayek, 1967, and Israel Kirzner, 1973; the subjectivist views of G. L. S. Shackle, 1969, 1972; etc.) have given the impression that genuine uncertainty and its effects cannot be represented with formal modeling tools. The approach suggested here is quite different: to harness the determinants of uncertainty in a modeling structure that characterizes regularity in behavior. Closely related ideas have also been recently analyzed by Richard Bookstaber and Joseph Langsam, 1983.

"optimal" behavior strategies, with little fruit in understanding observed behavior.

This trend is typified by recent Bayesian models of optimal risk behavior, which are synonymous with sophisticated continually updated response to new information. Some examples are optimal "search" models that specify various sequential strategies for job search, price or quality information, etc.[25] Yet, they bypass the issue that overides everything else: when to permit any search given the uncertainty in detecting whether the positive gains from efficient search strategies will outweigh the required search costs; especially when a diverse range of search opportunities might eventuate, and the timing of these future opportunities is also unknown.

## IV. Explaining Predictable Behavior: Framework and Illustration

The reliability theory briefly outlined above can be applied to the full spectrum of cases produced by different structures of uncertainty. It thus represents a general framework for analyzing behavior under all of these possibilities. On the other hand, standard choice theory analyzes the special case where there is no uncertainty due to a *C-D* gap.[26]

The narrowness of standard optimizing concepts is evidenced in the dominant tendency (even after years of extensive experience with conventional models) to steer away from incorporating genuine uncertainty into the analysis of behavior.[27] In contrast, the Reliability Condition directly harnesses the determinants of uncertainty to characterize regularity in behavior. This amounts to a reversal of the explanation assumed in standard economics, which places these determinants in the residual "error term" between observed behavior and the more systematic patterns claimed to result from optimization.

Thus, the idea of uncertainty as the source of predictable behavior is both a generalization and a major shift away from the explanatory framework of existing models, one that may be of importance to a number of fields. The following statements briefly summarize the major differences between the new framework (the economics of genuine uncertainty) and that of traditional optimization theory:

1) The basic theoretical objective is to understand the behavioral implications of genuine uncertainty, rather than the implications of maximizing for a given set of preferences or expectations. Genuine uncertainty results from a gap in an agent's decision competence relative to the difficulty in selecting more preferred alternatives, so that error and surprise cannot be avoided.

2) A wide range of factors contribute to uncertainty. In economics, these include cognitive limitations in processing given information or in interpreting potential information from the environment; vulnerable perceptions about preferences or expectations taken as given in traditional choice models; unreliable probability or expected utility information taken as given in standard risk-behavior theory. In addition, uncertainty may involve the ability to infer from past experience what was misunderstood that led to previous error; or the abil-

---

[25] See, for example, David Blackwell and M. A. Girshick, 1979; Thomas Ferguson, 1967; Peter Diamond and Michael Rothschild, 1978; Stephen Lippman and John McCall, 1979; Hirshleifer and John Riley, 1979.

[26] This conclusion also applies to the more recent models of behavior under uncertainty, which assume agents can infer reliable probabilities of future situations; and also recognize all possible events that might eventuate, or possible actions that might be useful to select. Such ability to comprehend the future is much more difficult than avoiding computational mistakes in a static world of known utility information over a fixed set of options. Consequently, these models are not moving closer but rather further away from dealing with genuine uncertainty due to a *C-D* gap. The reason is that in order to apply traditional optimizing concepts, the competence of the agent has been implicitly upgraded to handle the extra complexity resulting from an unpredictable future. On this issue, see the closing remarks of John Hey, 1979, pp. 232–34.

[27] For example, a recent statement by Lucas flatly concludes: "In situations of risk, the hypothesis of rational behavior on the part of agents will have usable content, so that behavior may be explainable in terms of economic theory.... In cases of uncertainty, economic reasoning will be of little value" (1981, p. 224; see also p. 223).

ity to identify potential actions which might be selected, or contingencies that might affect the consequences of future behavior.[28]

3) Optimizing with no uncertainty in choosing more preferred alternatives does not tend to produce systematic and stable regularity in behavior. Rather, it tends to destroy such regularity as successively more information can be reliably interpreted in guiding more complex behavior. This does not mean that formal optimization tools cannot be used, but rather that understanding how uncertainty affects behavior will systematically redirect the formulation of models and the questions to which they would be applied.

4) Predictable regularities of behavior are the manifestation of behavioral rules that represent patterns of behavior for which deviations exist that are preferred under certain conditions, but which are nevertheless ignored because of uncertainty in reliably interpreting potential information about when to deviate.

5) Intrinsic to behavioral rules is the ignoring or lack of alertness to potential information, the reaction to which would direct behavior into more complex deviations from such rules; even though such information may be costless to observe. Conversely, it is the alertness or sensitivity to information that determines the patterns and complexity of rules manifested in behavior. The Reliability Condition is a simple but general characterization of when greater flexibility to administer more complex behavior or to use more information will improve rather than worsen performance.

6) Behavioral rules not only involve outward symptoms of information sensitivity, but also internal mechanisms that generate such sensitivity. Thus, research in fields such as psychology, biology, and engineering has direct bearing on the structure of such rules. In contrast, traditional economic models have largely ignored research in these and other fields.

To help see the range of issues unified by the above analytical framework, a few illustrations are briefly presented.

### A. *The Consistency of Rule-Governed Behavior*

Traditional choice theory has tended to equate normative rationality with logical consistency of behavior, as described by various transitivity, intertemporal consistency, probability assessment, and other assumed conditions. For example, Jacques Drèze provides the following evaluation of the risk behavior axioms of standard expected utility theory:

> ... a consistent decision-maker is assumed always to be able to compare (transitively) the attractiveness of acts, or hypothetical acts and of consequences as well as the likelihood of events. These requirements are minimal, in the sense that no consistency of behaviour may be expected if any one of them is violated; *but they are very strong, in the sense that all kinds of comparisons are assumed possible, many of which may be quite remote from the range of experience of the decision-maker*. This is also the reason why the axioms have more normative appeal than descriptive realism; few people would insist on maintaining, consciously, choices that violate them, but their spontaneous behaviour may frequently fail to display such rigorous consistency.
>
> [1974, p. 11, emphasis added]

Drèze is quick to acknowledge and discount the descriptive validity of the expected utility axioms, but like many others he still feels secure in their normative validity in characterizing truly rational behavior under uncertainty.[29] Nevertheless, one might ask what would be the implication of a logically correct set of axioms (or a decision algorithm for search and learning behavior) *if obeying those axioms (or using the algorithm) would require the use and sensitive response to unreliable information* (for example, information remote from the range of experience of a decision maker)? To the extent this is the case, rule-governed behavior will ignore such axioms (or a decision algorithm) regardless of the logical properties violated in disobeying them.[30] Similar issues apply to traditional microeconomic theory. For example, what if preferences are less reliable for commodity bundles remote to a consumer's normal purchasing experience? Must we avoid this likely possibility in assuming fully connected preferences? Or is the violation of this assumption itself a major source of price-response regularities of consumers?

---

[28] The latter determinants have recently been described as particular types of uncertainty, such as parametric versus structural knowledge by Richard Langlois (1983) and "extended" uncertainty by Bookstaber and Langsam. They are extensions of the "unlistability problem" introduced by Shackle (1972). Whatever terminology or type of uncertainty is involved, we can characterize regularity in behavior depending on how each type of uncertainty affects the reliability of using information or selecting potential actions.

[29] See, for example, John von Neumann and Oskar Morgenstern, 1944, pp. 17–30; L. J. Savage, 1954, pp. 6–7, 19–21, 56–68, 82–84; Friedman and Savage, 1948.

### B. *Social Institutions Evolve Because of Uncertainty*

Neoclassical decision and general equilibrium models are typically without any explicit institutional structure, and have thus tended to direct attention away from questions about the evolution of particular forms of market organization and other social institutions.[31] In contrast, the Reliability Condition naturally suggests the systematic importance of such institutions to determine the scope and complexity of exchange relationships, and other social interactions involving cultural norms, customs, and aggressive behavior.

In this regard, it is noteworthy that Schotter's recent book on the theory of institutions defines them in a manner immediately implied by the Reliability Condition: "A social institution is a regularity in social behavior that...specifies behavior in specific recurrent situations, and is either self-policed or policed by some external authority" (p. 11).

Thus, evolved institutions are social rule-mechanisms for dealing with recurrent situations faced by agents in different societies. That is, institutions are regularities in the interaction between agents that arise because of uncertainty in deciphering the complex interdependencies created by these interactions.[32] I will return to this topic in Section V below, which considers the evolution of legal and exchange institutions.

A persistent theme in human literature illustrates a closely related issue that has been largely ignored by traditional choice theory; namely, the attempt of individuals to constrain or bind the flexibility of their actions.[33] A famous example in *The Odyssey* describes Ulysses trying to prevent himself from responding to the allurement of certain sirens: "...but you must bind me hard and fast, so that I cannot stir from the spot where you shall stand me...and if I beg you to release me, you must tighten and add to my bonds."

---

[30] See Paul Slovic and Amos Tversky, 1974, Slovic and Sarah Lichtenstein, 1983; Dirk Wendt, 1975; D. Conrath, 1973; Detlof Winterfeldt, 1980; and for systematic empirical evidence see Daniel Kahneman and Tversky, 1979, 1981, 1982; Ward Edwards, 1962; William Fellner, 1961; R. M. Hogarth, 1975. Some recent attempts to modify standard expected utility theory by dropping the "independence" or "substitution" axiom include Mark Machina, 1982 (equivalence relationships to global risk-aversion axioms); Graham Loomes and Robert Sugden, 1982, 1983 (regret theory); and S. H. Chew and K. R. MacCrimmon, 1979a, b; Peter Fishburn, 1981; R. Weber, 1982 (*alpha*-utility theory).

[31] See the following diverse range of analytical perspectives, including Alchian, 1950; James Buchanan, 1975, 1977; Buchanan and H. G. Brennan, 1981; Ronald Coase, 1937; Carl Dahlman, 1980; Demsetz, 1967; Hayek, 1967, 1973; Menger, 1871, 1883; Nelson and Sidney Winter, 1982; Andrew Schotter, 1981; Joseph Schumpeter, 1942; Oliver Williamson, 1975, 1979, 1981.

[32] Consider a person within a complex interdependent society, where uncertainty in deciphering these interdependencies quickly increases as they widen beyond his immediate experience. The Reliability Condition implies that his behavior will quickly become insensitive to nonlocal social contingencies. If among such contingencies are effects on other individuals, this implies a relatively sensitive or "self-interested" motivation toward a person's own self (and family), and away from alertness or "sympathy" toward other persons. This implication underlies the ideas Adam Smith developed in the *Theory of Moral Sentiments*, published prior to the *Wealth of Nations*. See Coase (1976) for a number of passages from the *Theory of Moral Sentiments*; for example, Smith, 1969, pp. 321–23, 347–48, 109–10.

[33] See John Elster, 1979; R. H. Strotz, 1955; and N. Howard, 1971. Another classic moral dilemma of great literature poses the protagonist in a situation with abnormally convincing information that "right circumstances" are at hand to engage in behavior precluded by social or religious norms.

### C. *Uncertainty and the Reliability of Expectations*

Both past and present economic models are crucially dependent on how they incorporate expectations in guiding behavior. Economists have been aware that beliefs about the future are often mistaken, and thus have been uneasy in both formulating and applying their models.[34] More recently, "rational expectations" models have attempted to resolve these problems by assuming that expectations correctly identify the mean and variance of stochastic variables that affect future environmental contingencies.[35] A key motivation for such models is to predict how "optimal" behavior will respond to changes in the structure of the environment, especially changes influenced by government policy. Yet, from a broader perspective, it is clear that most species that have evolved in nature exhibit relatively programmed behavior patterns that are highly insensitive to environmental changes, even if such rigidity results in their extinction. At best, such models could apply more broadly only by continually introducing specializing assumptions about the type of expectation "rationality" guiding the behavior of particular species.

Thus, in all of our existing models, either we are analyzing the maximizing response to possibly wrong expectations, or we avoid this issue by assuming expectations are reliable. In order to make progress in analyzing the role of expectations, we must understand how their use and formation are affected by genuine uncertainty in comprehending the future. For example, how reliable are agents' abilities to formulate beliefs about the future; and given the vulnerability of such beliefs, when will agents sensitively react to them, or when will they be alert to information that might prompt them to revise them?

### D. *The Pattern of Behavioral Complexity Evidenced in Nature*

My departure from standard choice theory was suggested by the general pattern of animals having a larger *C-D* gap than humans, yet regularity in their behavior is much more noticeable than for humans. The Reliability Condition implies a simple formal characterization of this overall pattern. Suppose we start with a given combination of the **p** and **e** variables, and consider a conceptual experiment where the **e** variables are held fixed, but the perceptual abilities of an agent are successively reduced compared to their initial effectiveness. This will increase the uncertainty in administering the initial behavioral repertoire, thus reducing the reliability ratios of particular actions. As already discussed, greater uncertainty will in general require a more inflexible structure of rules; that is, some of the actions in the initial repertoire must be excluded because their selection no longer satisfies the Reliability Condition.

Now apply this result to us as human observers watching other species with less reliable cognitive equipment than ourselves. We should notice a systematic pattern of greater rigidity and inflexibility in nonhuman species compared to our own behavior. This implication is testable to the extent that the effectiveness of different species' cognitive abilities can be independently measured from simply watching outward behavior (for example, relative brain to body mass). In addition, if we compare across a number of species, there should emerge a general pattern that correlates greater rigidity in behavior with less effective cognitive equipment.[36]

These implications characterize a pervasive qualitative pattern, one that is systematically evidenced in the comparative study of different species. Yet, they were obtained in a very simple way from the Reliability Condition. This is a significant indication that we are on the right track in understanding behavior, especially in developing a modeling structure that naturally suggests the very consideration of such questions.

---

[34] See John Hicks, 1935; Richard Muth, 1961; Axel Leijonhufvud, 1968, pp. 366–85, Rudiger Dornbusch and Stanley Fischer, 1978, pp. 270–75, 283–86.

[35] See for example, Thomas Sargent, 1979, and Lucas, 1981.

[36] For analysis of some of the more rigid, "forced" behavior movements of simple organisms, and other major instinctive patterns, see Roger Brown and Richard Herrnstein, 1975, pp. 23–31.

### E. *Explaining Instinctive Behavior*

The currently accepted explanation of instinctive rigidities is that they accomplish some function which is useful or adaptive most of the time for the natural environments in which they are exhibited.[37] But as already discussed, this feature is itself implied by the Reliability Condition; namely, that rule-governed behavior will ignore adjustment to unlikely contingencies, thus limiting response patterns to only the more probable or recurrent features of the environment. A number of implications concerning ecological structure, niches, extinction, etc., can also be derived (rather than simply described or assumed) from the analysis.

Explanation of specific behavioral rigidities can be obtained by using the Reliability Condition, $r/w > T$, with explicit variables and assumptions about an organism's perceptual components ($\mathbf{p}$) in terms of the sensory ($\mathbf{s}$) and cognitive ($\mathbf{c}$) attributes of particular organisms. In addition, we can introduce morphological ($\mathbf{m}$) attributes of organisms, along with the environmental variables ($\mathbf{e}$) which determine the structure of the environment. By understanding how these variables (denoted $\mathbf{z} = (\mathbf{s}, \mathbf{c}, \mathbf{m}, \mathbf{e})$) affect the reliability and tolerance limit components of the model, particular rule structures can be derived and compared with observed behavior of different organisms (including humans).

### F. *Brief Application to Imprinting*

Consider very briefly the phenomenon of *imprinting*.[38] Suppose that responding to a particular pattern in the environment is crucial to an organism's survival (for example, following its parents). Suppose also that without highly developed cognitive mechanisms, if the organism did not initially know the particular pattern, then it could not reliably distinguish that pattern from a number of similar patterns (i.e., a newly born organism could not reliably distinguish its parents from similar adults); but given a specific *reference pattern* to "lock onto," it can reliably distinguish it from other similar patterns. However, if the wrong pattern is locked onto, the organism's survival would be severely jeopardized.

In particular, the probability of right circumstances $\pi$ to lock onto a pattern is often a function of time since an organism's birth (for example, $\pi(t)$ is the chance of seeing only an organism's parents at time $t$ since birth). Recalling that the required reliability (i.e., the tolerance limit $T$) will quickly increase as $\pi$ drops to zero, we can derive the following two-stage behavioral process: stage one is a pattern-locking mechanism that reacts to whatever pattern first appears after the mechanism is initiated; while stage two is a resistance mechanism that severely constrains stage one to only certain sensitive periods for which the required reliability is very low (i.e., $\pi(t)$ is close to 1.0).

It can further be shown that the implied sensitive periods will be highly predictable across particular organisms of a species. In addition, comparative regularities across species in relatively sensitive learning periods can be derived. For example, we can characterize less rigidly patterned sensitivity phases in the development of human children in acquiring language, and the display of other cognitive skills.[39]

### G. *Punctuated Dynamics for Scientific Inquiry*

The work of Thomas Kuhn (1962) (see also Popper, 1969; Lakatos and Musgrave,

---

[37] The classic reference on instinct is Nino Tinbergen, 1951 (for example, pp. 151–84, especially 156–57 and 152–53). Other references include John Alcock, 1979, pp. 57–76, 87–102; Brown and Herrnstein, 1975, pp. 31–59; William Keeton, 1980, pp. 490–512, especially 503, 494, 496, 498; Eric Pianka, 1978, pp. 82–86, 152–53.

[38] See Alcock, 1979, pp. 67–73; Keeton, 1980, pp. 498–500; Konrad Lorenz, 1981, pp. 259, 275–87; David McFarland, 1982, pp. 303–05; W. R. Hess, 1973.

[39] See for example, Alcock, 1979, pp. 73–79; E. Mavis Hetherington and Ross Park, 1979; R. Grinder, 1962; Lawrence Kohlberg, 1966, 1969; N. Chomsky, 1972; J. Piaget, 1947, 1952. For related material from ethology, see Lorenz, 1981.

1970) has emphasized a systematic pattern of resistance in the behavior of scientists to quick and sensitive reaction to new ideas and theories. Yet, when sufficient anomalies and awkwardly interpreted evidence about a previous theory build up, a major shift in ideas (a "scientific revolution") will relatively quickly occur. This is an illustration of dynamic properties discussed below in Section VI. The Reliability Condition also implies other features in the behavior of scientists, such as: (a) resistance to accepting or using several competing theories unless there also exist easy to decipher (and reliable) criteria of when to switch between them; (b) similar resistance to incorporating new concepts or variables into accepted theories unless reliable criteria on how to use them are available (consider an economist's reaction to incorporating sociological variables into economic models); (c) differences in accepting and rewarding (salary, promotion, etc.) theoretical vs. empirical research in different fields depending on the reliability of observable data studied in those fields (for example, see Leijonhufvud's 1973 parody about "Life Among the Econ").[40]

### H. *Uncertainty and Consensus in Social Judgments*

Finally, in the area of ethics and social policy, consider the theory of justice advanced by John Rawls (1971). Underlying his whole analysis is the recognition that if individuals have reliable information about their own future circumstances (will they be smart or resourceful, or have special educational opportunities, or own highly valued property, etc.), they will respond to such information in the way they view social policies and institutions that would affect their particular situations.[41] This will produce a wide diversity of opinions about how to formulate and apply normative principles. Hence, in order to produce a highly uniform consensus or *regularity* in social judgments, Rawls introduced a pervasive uncertainty into the conceptual problem in the form of a "veil of ignorance." Such a procedure virtually eliminates reliable information (even in probabilistic form) about any particular individual's specific future circumstances that might eventuate depending on what principles are mutually agreed to by the whole group.[42] With a sufficient structure of uncertainty, individual judgments might be constrained to possibly a single, universally accepted principle of justice to guide social policy.

The important point is that the source of such a universal consensus, as well as the other behavior patterns discussed above, is uncertainty in using potential information about when to deviate from these regularities.

## V. Application to Economic Modeling

In this section, the Reliability Condition is briefly applied to a few modeling issues in standard economics.

### A. *Reluctance to Insure Against Rare Disasters*

Extensive empirical studies have shown that people are reluctant to insure themselves against large but rare disasters, in a manner that directly contradicts expected utility theory (see Howard Kunreuther et al., 1978). A recent statement by Kenneth Arrow summarizes the dilemma posed for standard "uncertainty" theory:

> A striking real life situation has given grounds for doubt about the validity of the expected utility hypothesis. Since 1969, the U.S. government has offered flood insurance at rates which are well below their actuarial value...Under the usual hypothesis of risk aversion, any

---

[40] Edward Leamer's work (1978) illustrates another issue about the reliability of model testing and formulation, which can be viewed as methodological rule-mechanisms to restrict "specification searches" used to claim empirical support for a theory. See Thomas Cooley and Stephen LeRoy (1981) for an application of Leamer's methodology to evaluating previous work on the demand for money.

[41] See Rawls, 1971, pp. 18–19, 137–38, 140, 149.

[42] See Rawls, 1971, pp. 150, 154–55. Notice in these passages how Rawls believes that a crucial feature of the veil of ignorance is the inability to formulate reliable probability information about the impact of social contingencies on particular individuals.

> individual should certainly be willing to undertake a favorable bet.... Yet, until the government increased the pressure by various incentives, very few took out this insurance.... The main distinguishing characteristic of those who took out flood insurance was acquaintance with others who took out insurance. This might be taken as an explanation in terms of information costs, but the information seems so easy to acquire and the stakes so large that this hypothesis hardly seems tenable.
> [1981, p. 2]

In contrast, the above analysis immediately suggests that even costless information will be ignored if the behavior resulting from its use will not satisfy the Reliability Condition, (recall that solving procedures for Rubic's cube systematically ignore costless information available simply by looking at the cube while unscrambling it). The real issue is why are agents reluctant to engage in behavior that might be prompted by such information.

Consider a brief sketch of the insurance behavior phenomenon. As the probability, $p$, of a disaster goes to zero, the number of such extremely rare but conceivable events grows indefinitely large. Given any positive setup costs of insuring against each of these possibilities, the total insurance cost will eventually exceed a person's (finite) wealth. Thus, it is clearly not appropriate to insure against all of them. (What do we call someone who is constantly trying to protect against rare but serious sickness; and what would happen to total output net of the demand for medical services if everyone exhibited this propensity?)

The above argument implies that the probability of the right time to insure, $\pi$, is bounded by the ratio of a person's wealth to total insurance cost; so that $\pi$ approaches zero as $p$ approaches zero. Thus, the required reliability will steeply rise for sufficiently rare disasters (i.e., the tolerance limit $T$ will begin to accelerate toward infinity as $p$ approaches zero—see Figure 1).[43] Note also that rare events are precisely those which are remote to a person's normal experience, so that uncertainty in detecting which rare disasters to insure against increases as $p$ approaches zero. Such greater uncertainty will reduce the reliability of insurance decisions (i.e., reduce the ration $r/w$) as disasters become increasingly remote to a person's normal experience.

As a result of the above factors, the required reliability of when to insure increases sharply just when the actual reliability is dropping. Thus, at some point as $p$ approaches zero, the Reliability Condition will be violated (i.e., $T$ will rise above the falling $r/w$ ratio). This implies people will switch from typically buying to typically ignoring insurance options, which is just the pattern documented in Kunreuther's 1978 study.

We can also show that after a person switches to ignoring insurance, he will be very reluctantly convinced to insure by any information source, *except those local to his normal experience* (for example, a neighbor, a relative, or an "acquaintance" as suggested in the above quotation). Note further that ignoring insurance does not necessarily mean a person consciously decides to ignore all the various potential insurance options—either those obtainable by contacting an insurance agent, or many other ones for which no market insurance is available.

This is a simple example of a more general implication: agents will only become alert or sensitive to information about options whose selection is reliable; or conversely, they will fail to become aware of information about options whose selection is unreliable. Another example of selective alertness to information is the use (or disuse) of marginal cost information to make production decisions, discussed next.

---

[43] Consider also very briefly the behavior of the $l/g$ ratio of the tolerance limit. The loss $l$ will be a negative function of the expected value of the disaster losses (denoted $E(p)$) relative to the expected value of the insurance costs (denoted $C(p)$); that is, $l$ is a negative function of $E(p)/C(p) = v(p)$, denoted $l(v(p))$. Similarly, $g$ is a positive function of $v(p)$, denoted $g(v(p))$. Now think of a sequence of actuarily fair or "pure" insurance options for which $v(p) \cong 1$. If the estimated $v(p)$ is close to zero, then the $l/g$ ratio will not deviate substantially from $l(1)/g(1)$. When this result is coupled with a steeply rising $(1-\pi(p))/\pi(p)$ ratio as $p \rightarrow 0$, we have the same acceleration implied for $T(p) \cong l(1)/g(1) \times (1-\pi(p))/\pi(p)$.

### B. *Spontaneous Alertness to Marginal Cost Information in Simple Production Environments*

A memorable episode in the history of economics was the marginalist controversy about whether businessmen use marginal cost calculations to guide their production decisions. The debate prompted Alchian to write "Uncertainty, Evolution, and Economic Theory" (1950). This was the article that first explicitly justified optimization theory as an explanatory tool to predict the outcome of selection processes (i.e., selection processes will produce simulations to optimizing behavior, which claim is contradicted by the above analysis). Regardless of how one views this debate, it is clear that businessmen typically do not use or are even aware of the kinds of marginal calculations discussed in standard production theory (this lack of awareness is itself an empirical regularity). But what would happen in a relatively simple production environment in which such information could be readily monitored and used with little uncertainty in directing production decisions?

The Reliability Condition implies the spontaneous development (without any special training in economic theory) of alertness and sensitive reaction to marginal cost information for sufficiently simple production environments. This will not be the usual situation, but are there cases that would naturally fit this hypothesis? An example is summarized in the following passage from Hirshleifer's price theory text:

> Electricity is typically generated by companies that operate a number of separate producing plants, with a transmission network providing connections to consumers as well as ties among the generating plants...the operating problem at any moment of time is to assign output most economically among the generating plants....
>
> Fred M. Westfield investigated the operating practices of a leading American electric utility. He discovered that this company employs a dispatcher to actually "assign the load" from moment to moment among the different plants. The dispatcher is guided by a Station-Loading Sliderule that shows what the economist would regard as the Marginal Cost function of each plant. By mechanically manipulating his Sliderule, the dispatcher automatically equates Marginal Cost for all plants in operation in such as way as to meet the total generation requirement....
>
> The company's method of division of output, and the Sliderule itself, were developed by engineers lacking the slightest acquaintance with economic theory. The company's engineers thus independently "discovered" Marginal Cost analysis.... [1980, pp. 286–87]

The engineers did not discover marginal cost analysis, but rather developed a way of reacting to what we as economists would call marginal cost information. Nevertheless, the development of a Station-Loading Sliderule is confirming evidence for the hypothesis of spontaneous sensitivity to marginal cost information in simple production environments. On the other hand, within standard price theory, it can only represent an isolated special case that illustrates a clearly noticeable use of such information.

### C. *Uncertainty Implies "Corridor" Dynamics for Macroeconomic Shocks*

A major issue in macroeconomic theory has a direct parallel with the insurance behavior phenomenon discussed above. Instead of deciding whether to insure against various natural disasters, think of a repertoire of activities to prepare for the negative effects of macroeconomic "shocks"; or more generally, anything that produces a coordination failure in an economic system.

When will an economic system evolve so as to "self-insure" against these potential sources of unemployment and other symptoms of coordination failure? Costly shock-preparation activities are beneficial if they are appropriately timed to mitigate the effects of a shock, but otherwise there is a loss from the reduction in output otherwise attainable.

Now suppose, analogous to the insurance case, that there are different types of shocks, some more severe than others; where larger shocks are possible but less and less likely to happen. In addition, the reliability of detecting when and how to prepare for large shocks decreases as their determinants and repercussions are more remote to agents' normal experience.

In a similar manner to that discussed for the insurance case, we can derive that the economy's structure will evolve so as to prepare for and react quickly to small shocks. However, outside of a certain zone or "corridor" around its long-run growth path, it will only very sluggishly react to sufficiently large, infrequent shocks. This is essentially the "corridor hypothesis" for macroeconomic systems recently advanced by Leijonhufvud (1981, pp. 103–29).

In this paper, I have not gone into the specific microprocesses involved (individual agent behavior, intra- and intermarket structures, transmission mechanisms, etc.). Nevertheless, even without adding more specific assumptions we can still derive this general qualitative feature as a necessary consequence of uncertainty. Standard economic theory has been unable to do so, as summarized by Leijonhufvud:

> ...general equilibrium theorists have at their command an impressive array of proven techniques for modelling systems that "always work well". Keynesian economists have experience with modelling systems that "never work". But, as yet, no one has the recipe for modelling systems that function pretty well most of the time but sometimes work very badly to coordinate economic activities. [1981, p. 103]

### D. *A Clear Prediction of the Law of Demand*

Suppose consumers do not have well-defined preference relations, but instead must deal with uncertainty in trying to detect when to buy more or less of particular commodities. Myriad "internal" perceptual and "external" environmental factors come together to determine the relative value of particular commodities. In a prospective sense, there is no reliable information to compare all the margins of choice to calculate the most preferred response for each future situation. Rather, consumers must try to react appropriately to various influences that might prompt them to purchase more of particular commodities.

Now suppose the price of a commodity $x$ rises. In order to benefit from continued purchases, the actual value of successive units of the commodity must exceed the now higher opportunity cost implied from the price increase. The likelihood of this situation arising is less than before, given the same structure of motivational influences affecting the value of $x$. Thus, the probability $\pi$ of the right situation to buy more $x$ is smaller. For the same reason, even when the right situation arises, the average excess of actual value over the higher price of $x$ (denoted $g$) is less than before. In addition, the average loss from purchasing more $x$ at the wrong time (denoted $l$) is now higher than before the price of $x$ went up. Each of these factors will increase the required reliability for purchasing $x$ (i.e., the tolerance limit for purchasing $x$, $T=(l/g)(1-\pi)/\pi$, will rise).

Given that $T$ has risen, *how is the consumer to change his behavior to be more reliable in purchasing* $x$? A general answer is suggested in an extensive literature in behavioral psychology about signal detection experiments.[44] The earliest experiments were similar to hearing tests where a person tries to detect the presence of a signal amid background noise (over a sequence of trials where the signal's occurrence is randomly distributed). A variety of other detection skills have been tested, which involve pattern recognition situations and various information processing and other cognative skills. All of the experiments exhibit a key feature: a person can increase the reliability of his detection behavior only by being more cautious in detecting the signals. That is, greater reliability requires a person to reduce the probability of reacting regardless of whether the signal is present or not. Note that reliability in these experiments is measured by the $r/w$ ratio used in the Reliability Condition (and reported in graphical form with *ROC* curves).[45]

---

[44] See David Green and John Swets, 1974; James Egan, 1975. A brief appendix on the signal detection experiments (plus some further material on reliability principles suggested by these experiments) is available on request from the author.

Now apply this principle to detecting when to buy more of a commodity $x$. A person can be more reliable in purchasing $x$ only by reducing the probability that potential influences will successively prompt him into purchasing (whether they be internal promptings, advertising, behavior of other consumers, or whatever).

Thus, we have a simple two-step syllogism: *a higher price requires purchasing behavior to be more reliable, which can be achieved only by reducing the probability of purchase*. This implication is essentially the law of demand for consumer behavior, yet without any qualification for income effects; nor must we use complicated Slutsky derivations, or other technical maximizing conditions. To some of us, the logic involved might even seem "too simple" compared to our intellectual investment in $n$-dimensional consumer theory. Nevertheless, in its simplicity is a clear, unambiguous implication of the Law of Demand, which we have never been able to derive with traditional optimizing methods.[46]

### E. *Evolution of Property Rights, Trading, and Market Structure*

Let me sketch a scenario about the evolution of an exchange system. Suppose initially that the reliable range of flexibility of agents' behavioral rules is more than sufficient to handle the complexity of the social environment (say in the primitive beginnings of human society). As a result, agent interactions evolve into more complex relationships in which the consequences from each agent's individual behavior depend on the actions of more and more other agents. In addition, the behavior of these other agents will become increasingly remote to the local experience of each agent as the network of social interdependencies broadens. Consequently, uncertainty in determining the consequences from selecting particular actions will successively increase for each agent in the society.

At some point, the evolution of more complex social interdependence will stop, unless social structures also evolve that reduce the scope of nonlocal information that individual agents must know to reliably forecast the consequences of their own behavior. (In more precise terms, the scope of information over which agents can reliably interpret successively narrows as the social environment becomes more complex.)

In general, further evolution toward social interdependence will require institutions that permit agents to know about successively smaller fractions of the larger social environment. That is, *institutions must evolve which enable each agent in the society to know less and less about the behavior of other agents and about the complex interdependencies generated by their interaction*.

One of the basic ways of accomplishing this is to divide up the decision authority to use resources so that only particular agents (or small groups of agents) have the right to control their use. With such a right-to-control institution, individual agents no longer have to know how other agents might use their "privately owned" resources. A whole range of factors that are within an agent's local experience can now be used to determine the consequences of particular use decisions. Two of the more important possibilities are decisions about whether to consume or delay the use of a resource, and about whether to transfer the right-to-control resources to other agents. Obtaining the right to control itself becomes valuable, given that only local information is now required to control the use of a resource.

---

[45] *ROC* stands for "receiver operating characteristic"; see Green and Swets, pp. 31–34.

[46] The Reliability Condition also implies a number of other key empirical regularities that are not derivable from basic maximization theory (see fn. 5 above). Another implication is that behavior will be relatively sensitive to information that defines an agent's local frame of reference within the environment. This will produce "framing effects" studied by Kahneman and Tversky, and a number of other anomalies now widely recognized in the risk-behavior literature (see fn. 30 above). Still other examples include the "excessive reaction" of securities and futures markets to "current information"; the "tendency to ignore prior information" used in Bayesian probabilities; and the "insensitivity of judgments to sample size"; even by "professionally trained" econometricians (see Arrow, 1981, pp. 3–7).

In more basic terms, the question is whether agents will be willing to cooperate with each other through increasingly complex interdependencies that have the potential—if properly coordinated—to increase average output per agent. As the society becomes more complex, agents will cooperate only in ways that enable them to use increasingly local information to detect whether they will individually benefit. That is, they will exhibit a "propensity to cooperate" only in situations where increasingly local experience indicates a benefit—even if such restriction cuts off a whole range of benefits that might result from more subtly interconnected forms of cooperation. A major way of satisfying this restrictive criteria is to cooperate only in situations where agents immediately reciprocate the cooperative actions of each other, such that each perceives a net benefit based on his own self-evaluation of the forsaken and received items.

This form of reciprocation enables agents to decide based on immediately local experience about the results from cooperating. Thus, their tendency to cooperate in such situations will be relatively great compared to myriad other possibilities that would require the reliable use of more nonlocal information to avoid mistakes. (In more precise terms, we can show that the probability of agents cooperating in such situations will be much higher than for other forms of cooperation.) This limited tendency to cooperate can itself be regarded as a behavioral regularity, one that Adam Smith recognized as the "propensity to truck or barter." Notice also that such a propensity depends on a structure of property rights that enables agents' self-evaluations to determine the use of resources without knowing the behavior of other agents.

The above discussion is only a brief illustration of a large number of implications about legal and market institutions. These institutions will evolve so as to provide predictable opportunity for mutual reciprocation situations; and so as to reduce the scope and complexity of information that must be reliably interpreted for agents to benefit from these situations. For example, a few implications include: a restriction to more centralized market organization and to financial instruments that enable agents to avoid knowing the particular circumstances, attributes, and identity of potential reciprocators and the items reciprocated; a severe restriction of futures markets and auction markets to certain strategic locations within a larger network of inventory markets structured so as to reduce price fluctuations;[47] and ownership structures that enable agents to avoid detecting whether continued reciprocation will be maintained, especially when this is necessary for particular reciprocators to realize longer term benefits or to prevent certain losses.[48] The essential factor in all of these institutional regularities is uncertainty in deciphering the complexity of the social environment.[49]

Finally, let me mention another key feature about the possibility of coordination failures. A complex cooperative system must somehow limit the occurrence of serious coordination failures. Nevertheless, its very complexity can evolve only to the extent that it enables agents to benefit without deciphering more than a tiny fraction of its overall structure. As a result, a complex system cannot prevent coordination failures that would require agents to understand a sizeable fraction of its complexity in order to avert them.

---

[47] A few modern references on the above topics are: Alchian, 1969, 1977; Robert Clower, 1967; Clower and Leijonhufvud, 1975; Robert Jones, 1976; Seiichi Kawasaki et al., 1982; Lester Telser, 1981.

[48] The reliability model can be used both to predict the likelihood of opportunistic behavior (discontinuing reciprocation), and how the likelihood of such behavior affects the required reliability of various kinds of contractual arrangements. In many cases, the only solution is to structure ownership of assets in a way that eliminates having to detect when to engage in certain contracts. This will produce a stable regularity in contractual and market ownership patterns, which are also studied under the rubric of "transaction costs" (see Williamson, 1975, 1979, 1981, 1983; Benjamin Klein et al., 1978; also Alchian and Demsetz, 1972; Coase, 1937, 1960; Demsetz, 1969, 1967; Dahlman, 1980).

[49] Standard choice theory concentrates exclusively on the potential gains from trade (via Edgeworth exchange boxes, etc.), rather than on the effect of uncertainties created in trying to realize that potential. Consequently, we now have an elaborate general equilibrium theory of exchange which is devoid of the very institutional regularities necessary for complex exchange economies to evolve in the first place (see the epilogue of Vernon Smith, 1982, p. 952).

## VI. Switching and Punctuation Dynamics

Recall the notation introduced above in Section IV, Part E, where $\mathbf{z} = (\mathbf{s}, \mathbf{c}, \mathbf{m}, \mathbf{e})$ represents an agent's sensory (**s**), cognitive (**c**), and morphological (**m**) components (hereafter denoted by $\mathbf{y} = (\mathbf{s}, \mathbf{c}, \mathbf{m})$), along with the environmental variables **e**. Using these variables, we can analyze how uncertainty affects the dynamic response of behavioral rules, and how agents' internal components interact with each other and with the environment to generate evolutionary change in themselves and in the surrounding environment. Two key dynamic properties are conditioned by the transition point between satisfying or violating the Reliability Condition (i.e., the point at which $r/w = T$).

*First*: Changes in the environmental variables **e** may shift the reliability ratio $r/w$ or the tolerance limit $T$ of an action; causing them to "cross over" each other from their initial positions (i.e., shift $r/w$ from below $T$ to above it, or vice versa). If this happens, rule-governed behavior will switch from allowing to severely restricting that action. Thus, a relatively sudden "switching" between different behavior patterns may occur.

*Second*: If the reliability ratio of an action is initially bounded below its tolerance limit, then behavioral rules will prohibit that action. Now consider a small change in a particular component, $y^0 \in \mathbf{y}$, which would shift $r/w$ and $T$ for such an action closer together, but not enough for them to cross over each other. So long as this is the case, there will be no change in an agent's behavior that might improve or worsen his performance, because the Reliability Condition for selecting that action is still violated. Suppose, however, that movement in some of the *other* **z** variables besides $y^0$ (which might include the **e** variables) shift $r/w$ and $T$ sufficiently for them to cross over each other.

At the point of transition, greater reliability from changes in $y^0$ will now allow selecting the action to improve an agent's performance; which may initiate evolutionary adjustment of $y^0$ in the appropriate direction. This means that the **y** attributes may exhibit relatively sudden increases or decreases in the speed of evolutionary change. Thus, evolutionary adjustment in the **y** attributes may be "punctuated" with a variety of sudden changes, especially as a large number of such attributes interact through an agent's behavioral rules, or the environment is itself influenced by the actions of other agents.

It is significant that a simple "crossover" mechanism will generate irregular dynamic movement in the outward behavior or internal attributes of an agent, and suggests an alternative to the recent attempts to account for such effects via catastrophe theory.[50] A recursive use of the Reliability Condition can also generate systematic hysteresis effects, in which the crossover point depends on the past history and direction of a variable's movement.[51]

A few examples to illustrate the above two dynamic properties are the following:

1) A number of implications characterize sudden switching of animal behavior between different actions, such as aggressive behavior in either attacking or retreating, or territorial behavior in either attack or defense strategies. A common example in economics involves switching between buying and selling strategies in financial markets, resulting in sudden movement in stock prices. In general, a wide range of behavior in economics is governed by such switching and hysteresis effects and has been obscured by the use of traditional optimization theory.

2) A specific economic illustration of the crossover mechanism is the "corridor hypothesis" for macroeconomic systems discussed above in Section V. Another example

---

[50] See E. C. Zeeman, 1977; Rene Thòm, 1975; David Berlinski, 1975; Hector Sussman, 1975.

[51] Consider very briefly a two-stage use of the Reliability Condition. First, $r/w = T$ is used to characterize a transition point between different behavior patterns. Second, introduce uncertainty about an agent trying to detect unstable shifts in this transition point. An agent may fail to switch once the transition point is reached, or he might mistakenly switch too early. A second application of the Reliability Condition implies that an agent will delay switching until he rarely switches too early; so that the observed switching point will shift depending on the action selected before the switch occurred.

is a structure of expectation "stages" during inflations (ranging from initially "sluggish" to eventually "explosive" expectation adjustment), which contrasts with recent rational expectations modeling.[52]

3) Growing evidence supports the "punctuation hypothesis" recently advanced in evolutionary biology, which claims that irregular bursts in the pace of evolutionary change have produced speciation and macroevolution of dramatic morphological changes (see Stephen Gould and Niles Eldredge, 1977; Steven Stanley, 1979).

4) An example of the latter which has been of considerable interest is the dramatic expansion of the cerebrum responsible for the higher thought processes of humans. Of all the various y attributes, the cerebrum most directly tends to prompt increasingly sophisticated behavior patterns. Right situations for selecting particular actions within a behavioral repertoire will become increasingly rare as the complexity of that repertoire increases.[53] This will cause a steeply rising acceleration in the required reliability for selecting these actions. Thus, if the second dynamic property above ever triggers rapid expansion of the cerebrum, then its sudden leveling off at a larger size is also implied. This dynamic pattern has been of interest and puzzlement in the biology literature.[54]

## VII. Hierarchical Structure and Evolution of Reliable Complexity

We can also characterize how uncertainty may generate hierarchical structures of increasingly flexible rules. Such rule-hierarchies have far reaching applications, some of which are briefly discussed in the following remarks:

1) For example, consider a system of components that interact with each other at level $v$, while these interactions comprise a larger system that interacts within a surrounding environment at level $v+1$. For simplicity, the relationship between a system and its subcomponents is functionally written, $s_{v+1}(\mathbf{s}_v)$, where $s_{v+1}$ denotes the system and $\mathbf{s}_v$ denotes its subcomponents. Thus, we have a recursive structure of rule-governed systems, $s_{v+1}(\mathbf{s}_v)$, where each element of $\mathbf{s}_v$ is itself a system of components at the next lower level $v-1$, denoted $s_v(\mathbf{s}_{v-1})$.

2) Now, suppose that more simply structured subcomponents decrease the reliability of a system in administering more complex interactions with its environment. For example, such components might be more vulnerable in distinguishing nonlocal phenomena. For any given level of subcomponent structure, viable performance requires a minimum degree of behavioral rigidity. Thus no system composed of similarly structured components can allow greater flexibility without hindering its viability. Consequently, the only way more sophisticated behavior could arise from such systems is for a number of them to evolve into the subcomponents of a still larger system. Since the components of the larger system are recursively built up from smaller subsystems, additional structure may be permitted which enables them to reliably guide more complex behavior of a larger system. When this is possible, the behavior of the larger system can be less rigidly constrained than its component subsystems.

3) Recent discoveries in microbiology dramatically illustrate this implication. They show how molecular mechanisms direct the embryological unfolding of living systems. The essential feature of all of these mechanisms are large molecular structures (containing hundreds or thousands of atoms) that interact with each other literally by recognizing each other's shape. That is, they interact with noncovalent bonds which are very much weaker than the covalent bonds (i.e., the merging of electron clouds) of physical chemistry. Thus, stable bonding requires a relatively large surface closely matched to

---

[52] The explosive stage could refer to the final phase of a hyperinflation in which agent's expectations so quickly adjust that trying to counteract this reaction by further money supply acceleration will drive real balances toward zero.

[53] As an agent's behavior becomes more complex, each additional action must compete against more and more other actions. Thus, the likelihood of an additional action being more preferred than other actions is conditional on the behavioral complexity of an agent; and in general will decrease as the complexity of his repertoire increases.

[54] See, for example, Edward Wilson, 1975, pp. 547–50; Jerison, 1973, pp. 402–43; D. Pilbean, 1972.

the shape of another molecule (which large surfaces require many atoms within each molecule).[55] Molecular shape enables the precise calibration of "stereo-specific" bonding, which permits a much more complex structure of interaction possibilities than otherwise possible with the more rigid constraints of physical chemistry. Moreover, the precise recognition properties of stereo-specific bonding enables the reliable direction of complicated molecular mechanisms, as evidenced in the biochemistry of cell regulation and embryological processes.[56] The significance of this is summarized by Nobel Prize biochemist Jacques Monod:[57]

> [Stereo-specific bonding gave] molecular evolution a practically limitless field for exploration and experiment, [which] enabled it to elaborate the huge network of cybernetic interconnections which makes each organism an autonomous functional unit, whose performances appear to transcend the laws of chemistry if not to ignore them altogether. [1972, p. 78]

4) If the recursive structure, $s_{v+1}(\mathbf{s}_v)$, is continued to higher or lower levels ($v+2, v+3,\ldots$; or $v-1, v-2,\ldots$) we obtain a hierarchical structure of increasingly sophisticated systems, where later stages are governed by successively more flexible rules. Such hierarchical structures represent a basic way systems conditioned by uncertainty can evolve into allowing successively more sophisticated behavior without hindering their viability in the process.

This pattern of hierarchical development is systematically evidenced in nature at a number of intertwining levels. For example, there are the invariable behavior patterns of atoms, which are composed of successively more basic subatomic particles; and which are themselves components of larger cosmological systems whose behavior is also synonymous with highly predictable laws.[58] Above this level, there is another hierarchy of organic molecules (discussed above) that eventually form components in living cells. Such cells in turn are subcomponents of still larger organs and tissues that permit relatively more flexible behavior of yet another hierarchy of increasingly sophisticated living organisms. Finally, there is the subtle, usually difficult to predict, behavior of humans and their social institutions. Looking back on this structure, the particular course of its evolution may be extremely improbable. Nevertheless, what did evolve has been through a hierarchical process from very predictable to relatively much less predictable phenomena.

5) Hierarchical structures may also have systematic importance in the design of cognitive and related (natural and artificial) learning processes.[59] For example, there may be

---

[55] See J. Monod, 1972, pp. 45–46.

[56] See J. Monod, 1972, chs. 4–7.

[57] For a recent more technical overview of the subject, see James Watson, 1976.

[58] The body of this paper has only briefly alluded to the physical sciences. At issue is whether the invariable regularities exhibited by natural phenomena can be regarded as "rule-mechanisms" to cope with extreme uncertainty in avoiding destabilizing interactions between the components of a system that might disintegrate its structure? More generally, what patterns of component interaction are viable in the sense of generating their own continuation, or the continuation of larger interactive patterns between components which are themselves systems? Many topics in the physical sciences could be discussed, but only three topics are mentioned here. First, we can analyze uncertainty in producing stable macrostructures to characterize relationships between the "particles" of matter and the "forces" that interconnect them. Second, we can consider uncertainty in maintaining the structural stability of tightly compacted systems to characterize symmetry properties, and other statistical regularities studied in quantum mechanics. Third, we can analyze the effects of violating the general relativity postulate of modern physics, especially about uncertainties in dealing with complex interdependencies permitted without the constraint of generally covariant interactions. On these three topics see respectively: P. C. W. Davies, 1979; J. P. Elliot and P. G. Dawber, 1979; Enrico Cantore, 1969; Albert Einstein, 1952, 1956.

Underlying these regularities is a persistent theme about the unity of science, as suggested in the following remark by Einstein: "The most incomprehensible thing about the universe is why it is so comprehensible." (See also the closing remarks of Kuhn's 1962 essay, p. 173.) The answer may lie in how extreme uncertainty affects the structure of self-continuing physical systems.

[59] See Simon, 1969, 1979a; Newell and Simon, 1972; J. R. Anderson, 1980; G. T. Miller et al., 1960.

resistance to knowledge not built up in recursive stages. For example, explicitly hierarchical methods have evolved in the above-mentioned strategy books on playing blackjack, and more recently on how to solve Rubik's cube.[60]

6) It has also been argued by Kohlberg, with extensive supporting experimentation, that the moral development of children as they mature into adults follows a highly patterned hierarchical structure of six stages. The first stage is guided by "blind obedience to rules and authority...," which proceeds though intermediate steps to stage six, which is "guided by self-chosen ethical principles."[61] A pattern of successively more complex moral judgments is clearly suggested in this hierarchy.

7) The viability of an evolving system, (for example an ecological system of organisms, or an exchange system of competing agents), which originates truly novel change (whose interactive possibilities are largely unrelated to the system's past history) may be quite sensitive to uncertainty in avoiding disruptive novelty. If this is the case, the very processes which generate and select such novelty will themselves be organized in a hierarchical structure of increasingly flexible rule-mechanisms. An important illustration is the structure of relationships that connect the rigidly patterned molecular design of DNA to the more visible interactions comprising natural selection.[62]

Other implications characterize the diversity and pace of novel change that can be reliably controlled by an evolving system. For example, a more rapid average pace is permitted as the reliability of selective processes increases. These implications underly the major differences in the qualitative nature and average speed of cultural compared to biological evolution.

## VIII. Conclusion

I have argued that uncertainty is the basic source of predictable behavior, and also the main conditioning factor of evolutionary processes through which such behavior evolves. Uncertainty exists because agents cannot decipher all of the complexity of the decision problems they face, which literally prevents them from selecting most preferred alternatives. Consequently, the flexibility of behavior to react to information is constrained to smaller behavioral repertoires that can be reliably administered. Numerous deviations from the resulting behavior patterns are actually superior in certain situations, but they are still ignored because of uncertainty about when to deviate from these regularities.

In contrast, standard economics analyzes the special case of no uncertainty in selecting most preferred options. This way of understanding behavior forces the determinants of uncertainty into the residual "error term" between observed behavior and the more systematic patterns claimed to result from optimization. I am thus suggesting a reversal of the explanation assumed in standard economics: the factors that standard theory places in the error term are in fact what is producing behavioral regularities, while optimizing will tend to produce sophisticated deviations from these patterns. Hence, the observed regularities that economics has tried to explain on the basis of optimization would disappear if agents could actually maximize.

---

[60] A good example of a hierarchical method is Kosniowski, 1981, especially in contrast to David Singmaster, 1979 (called the "definitive treatise" by *Scientific American*), which follows a complex, cyclical development of ideas that switches back and forth between different parts of the book. Singmaster's book is also several times longer than later books (cited above in fn. 14), both in terms of number of words and notational density.

[61] See Kohlberg, 1976, pp. 30 and 32; and 1963, 1969.

[62] Consider the extreme uncertainty of tiny molecular structures directing the construction of living systems. Maurice Wilkin's 1953 paper (which accompanied Watson and Cricks' original paper in *Nature*) begins: "While the biological properties of deoxypentose nucleic acid suggest a molecular structure containing great complexity, X-ray diffraction studies described here show the basic molecular configuration has great simplicity" (p. 738). The Reliability Condition implies the opposite presumption; namely, that precisely because DNA is the ultimate source of larger biological systems, whose complexity cannot be reliably manipulated from any interaction local to its tiny structure, its internal design must be both rigidly patterned and engineered to replicate virtually without guidance from its local chemical environment.

Another basic conclusion is that appropriately structured behavioral rules will not necessarily arise. Rather, they will evolve to the extent that selection processes quickly eliminate poorly administered behavior. This will more likely occur when agents are involved in highly competitive interactions that themselves indirectly result from scarcity. However, if weak selection processes are present, relatively vulnerable or dysfunctional behavior may evolve.

One area of major normative significance is the development of human social institutions; in particular, political institutions that have the opportunity to influence the outcomes generated by exchange competition. This is especially important if human agents are able to foresee numerous potential cases where the cooperative results of exchange institutions could be improved, but without being able to reliably administer the additional complexity necessary to realize those improvements.

Think of this issue in terms of the Reliability Condition. People may be able to identify government actions where situations exist in which a society will benefit (i.e., the probability of right circumstances $\pi$ for selecting these actions is positive). Nevertheless, they may be unable to administer these actions with sufficient reliability to benefit the society by adding them to the government's repertoire of authorized activities (i.e., $r/w<T$ even though $\pi>0$). If this is the case, the society will benefit by appropriately limiting the scope and complexity of government behavior.

But how is such limitation to arise? It is here that we enter the area of "constitutionalism," defined broadly as the design of rule-mechanisms to restrict the flexibility of government to react to whatever influences might prompt it to engage in vulnerable activities. The writings of seventeenth- and eighteenth-century political philosophers and statesmen were primarily concerned with these issues. Out of their efforts came a number of features incorporated in the United States Constitution, such as the separation of powers mechanism.[63]

On a wider scale, the history of civilization can be organized around a theme of groping for social rule-mechanisms.[64] Nevertheless, the understanding of such mechanisms is only in its rudimentary beginnings; and in the last hundred years, the general trend has been away from these topics—especially for analysts trained in mainstream economic theory.[65] The reason is that mainstream theories have systematically directed attention away from the study of processes that limit flexibility to choose potentially preferred actions. A refocusing of research on such processes—with the appropriate analytical framework to guide us—may have practical consequences for the viability of existing institutions.

---

[63] The basic source materials on these issues are the Federalist Papers by Hamilton and Jefferson (for example, numbers 10, 47, 48, 51). For a modern reference, see Martin Diamond, 1981.

[64] The often seemingly bizarre practices of religion and cultural ritual may also represent the design technologies of social rules crucial to the coordination and intensification of social bonds. For some interesting readings about ritual, symbolism, and comparative religion, see William Lessa and Evon Vogt, 1979, and M. Gluckman, 1962.

[65] Notable exceptions to this general trend are the writings of Buchanan and Hayek (see fn. 31 above).

## REFERENCES

**Abernathy, William J.**, *The Productivity Dilemma: Roadblock to Innovation in the Automobile Industry*, Baltimore: Johns Hopkins University Press, 1978.

______, **Clark, Kim and Kantrow, A.**, "The New Industrial Competition," *Harvard Business Review*, September-October 1981, *59*, 68–81.

**Alchian, Armen A.**, "Uncertainty, Evolution and Economic Theory," *Journal of Political Economy*, June 1950, *58*, 211–21.

______, "Information Costs, Pricing and Resource Unemployment," *Western Economic Journal*, June 1969, *7*, 109–28.

______, "Why Money?," *Journal of Money,* *Credit, and Banking*, February 1977, *9*, 133–40.

______ **and Demsetz, Harold,** "Production, Information Costs, and Economic Efficiency," *American Economic Review*, December 1972, *62*, 777–95.

**Alcock, John,** *Animal Behavior: An Evolutionary Approach*, Sunderland: Sinauer Associates, 1979.

**Aldrich, R.,** *Organization and Environments*, Englewood Cliffs: Prentice-Hall, 1979.

**Alexander, Richard,** *Darwinism and Public Affairs*, Seattle: University of Washington Press, 1979.

______ **and Borgin, G.,** "Group Selection, Altruism, and the Levels of Organization of Life," *Annual Review of Ecology and Systematics*, September 1978, *9*, 449–75.

**Anderson, Ian,** *Turning the Tables on Las Vegas*, New York: Harper & Row, 1975.

**Anderson, J. R.,** *Cognitive Psychology and Its Implications*, San Francisco: W. H. Freeman, 1980.

**Archer, John,** *The Archer Method of Winning at 21*, Hollywood: Wilshire Book Company, 1973.

**Arrow, Kenneth J.,** "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1981, *20*, 1–9.

______, "Vertical Integration and Communication," *Bell Journal of Economics*, Spring 1975, *6*, 173–83.

**Ashby, W.,** *An Introduction to Cybernetics*, New York: Wiley, 1956.

**Athos, Anthony G. and Pascale, Richard T.,** *The Art of Japanese Management*, New York: Simon & Schuster, 1981.

**Axelrod, Robert,** (1980a) "Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution*, March 1980, *24*, 3–25.

______, (1980b) "More Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution*, September 1980, *24*, 379–403.

**Barlow, B. W.,** *The Cube: A Short and Easy Solution*, Salt Lake City: Hawkes Publishing, 1981.

**Berlinski, David,** "Mathematical Models of the World," *Synthese*, August 1975, *31*, 211–27.

**Black, M. R. and Taylor, H.,** *Unscrambling the Cube*, Burbank: Zephyr Engineering Design, 1980.

**Blackwell, David and Girshick, M. A.,** *Theory of Games and Statistical Decisions*, New York: Dover, 1979.

**Bookstaber, Richard and Langsam, Joseph,** "Coarse Behavior and Extended Uncertainty," Working Paper 83–1, Graduate School of Management, Brigham Young University, 1983.

**Boland, Lawrence, A.,** "On the Futility of Criticizing the Neoclassical Maximization Hypothesis," *American Economic Review*, December 1981, *71*, 1031–36.

**Bradshaw, C. M., Szabadi, E. and Bevan, R.,** "Behavior of Humans in Variable-Interval Schedules of Reinforcement," *Journal of the Experimental Analysis of Behavior*, September 1976, *26*, 135–41.

**Brown, Roger and Herrnstein Richard,** *Psychology*, Boston: Little-Brown, 1975.

**Buchanan, James M.,** *Freedom in Constitutional Contract*, College Station: Texas A&M University Press, 1977.

______, *The Limits of Liberty: Between Anarchy and the Leviathan*, Chicago: University of Chicago Press, 1975.

______ **and Brennan H. G.,** *Monopoly in Money and Inflation*, London: Institute for Economic Affairs, 1981.

**Buskist, William F. and Miller, Harold L., Jr.,** "Concurrent Operant Performance in Humans: Matching When Food is the Reinforcer," *Psychological Record*, January 1981, *31*, 95–100.

**Bussert, Patrick,** *You Can Do the Cube*, New York: Puffin Books, 1981.

**Cagan, Phillip,** "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956.

**Caldwell, B.,** *Beyond Positivism: Economic Methodology In the Twentieth Century*, London: Allen & Unwin, 1982.

**Campbell, Donald C.,** "On the Conflicts Between Biological and Social Evolution and Between Psychology and Moral Tradition," *American Psychologist*, December 1975, *30*, 1103–26.

______, "Downward Causation in Hierarchically Organized Biological Systems," in F. J. Ayala and T. Dobzhausky, eds., *Studies in the Philosophy of Biology*, New York: Macmillan, 1974.

**Canfield, Richard A.,** *Blackjack: Your Way to Riches*, Secaucus: Lyle Stuart, Inc., 1979.

**Cantore, Enrico,** *Atomic Order: An Introduction to the Philosophy of Microphysics*, Cambridge: MIT Press, 1969.

**Cardoza, Avery D.,** *Winning Casino Blackjack for the Noncounter*, Santa Cruz: Cardoza School of Blackjack, 1981.

**Carter, C. F. and Williams, B. R.,** *Industry and Technical Progress: Factors Governing the Speed of Application of Science*, New York: Oxford University, 1957.

**Cavelli-Sforza, L., and Feldman, Marcus,** *Cultural Transmission & Evolution: A Quantitative Approach*, Princeton: Princeton University, 1981.

**Chandler, Alfred D.,** *Strategy and Structure: Chapters in the History of Industrial Enterprise*, Cambridge: MIT Press, 1962.

______, *The Visible Hand: The Managerial Revolution in American Business*, Cambridge: Belknap Press, 1977.

**Cheung, Steven,** "Transactions Costs, Risk Aversion, and the Choice of Contractual Arrangements," *Journal of Law and Economics*, April 1969, *12*, 23–42.

**Chew, S. H. and MacCrimmon, K. R.,** (1979a) "Alpha-Nu Choice Theory: A Generalization of Expected Utility Theory," Working Paper No. 669, University of British Columbia, 1979.

______ **and** ______, (1979b) "Alpha Utility Theory, Lottery Composition, and the Allais Paradox," Working Paper No. 686, University of British Columbia, 1979.

**Chomsky, N.,** *Language and Mind*, New York: Harcourt, Brace & Jovanovich, 1972.

**Clark, R.,** *The Japanese Company*, New Haven: Yale University Press, 1979.

**Clower, Robert W.,** "The Keynesian Counterrevolution: A Theoretical Appraisal," in F. Hahn and F. Brechling eds., *The Theory of Interest Rates*, London: Macmillan, 1965.

______, "A Reconsideration of the Microfoundations of Monetary Theory," *Western Economic Journal*, December 1967, *6*, 1–8.

______ **and Leijonhufvud, Axel,** "The Coordination of Economic Activities: A Keynesian Perspective," *American Economic Review Proceedings*, May 1975, *65*, 182–88.

**Coase, Ronald H.,** "The Nature of the Firm," *Economica*, November 1937, *4*, 386–405.

______, "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

______, "Adam Smith's View of Man," *Journal of Political Economy*, October 1976, *19*, 529–46.

**Cohen, Michael D. and Axelrod, Robert,** "Coping with Complexity: The Adaptive Value of Changing Utility," *American Economic Review*, forthcoming.

**Conner, M.,** *The Tangled Wing: Biological Constraints on the Human Spirit*, New York: Basic Books, 1982.

**Conrath, D.,** "From Statistical Decision Theory to Practice: Some Problems with the Transition," *Management Science*, April 1973, *19*, 873–94.

**Cooley, Thomas F. and LeRoy, Stephen F.,** "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

**Cooter, Robert,** (1982a) "The Cost of Coase," *Journal of Legal Studies*, January 1982, *11*, 1–34.

______, **Marks, Stephen and Mnookin, Robert,** (1982b) "Bargaining in the Shadow of the Law: A Testable Model of Strategic Behavior," *Journal of Legal Studies*, June 1982, *11*, 225–52.

**Dahlman, Carl,** *The Open Field System and Beyond*, Cambridge: Cambridge University Press, 1980.

**Davies, P. C. W.,** *The Forces of Nature*, Cambridge: Cambridge University Press, 1979.

**Demsetz, Harold,** "Information and Efficiency: Another Viewpoint," *Journal of Law and Economics*, April 1969, *12*, 1–22.

______, "Toward a Theory of Property Rights," *American Economic Review Proceedings*, May 1967, *57*, 347–59.

**de Villiers, P.,** "Choice in Concurrent Schedules and a Quantitative Formulation of the Law of Effect," in W. K. Honig and J. E. R. Standdon, eds., *Handbook of Operant Behavior*, Englewood Cliffs: Prentice-Hall, 1977, 233–87.

**Diamond, Martin,** *The Founding of the Democratic Republic*, Itasca: Peacock, 1981.

**Diamond, Peter and Rothschild, Michael,** *Uncertainty in Economics*, New York: Academic Press, 1978.

**Dornbusch, Rudiger and Fischer, Stanley,** *Macro-Economics*, New York: McGraw-Hill, 1978.

**Drèze, Jacques H.,** "Axiomatic Theories of Choice, Cardinal Utility, and Subjective Probability," a review in his *Allocation Under Uncertainty: Equilibrium and Optimality*, New York: Wiley, 1974, 1–23; reprinted in P. Diamond and M. Rothschild, eds., *Uncertainty in Economics*, New York: Academic Press, 1978, 37–57.

**Dubey, Leon B., Jr.,** *No Need to Count: A Practical Approach to Casino Blackjack*, San Diego: A.S. Barnes & Co., 1980.

**Edwards, Ward,** "Subjective Probabilities Inferred from Decisions," *Psychological Review*, March 1962, *69*, 109–35.

**Egan, James P.,** *Signal Detection Theory and ROC Analysis*, New York: Academic Press, 1975.

**Einstein, Albert,** *The Meaning of Relativity: Including the Relativistic Theory of the Non-Symmetric Field*, Princeton: Princeton University Press, 1956.

______, *The Principle of Relativity*, New York: Dover, 1952.

**Elliot, J. P. and Dawber, P. G.,** *Symmetry in Physics*, Vols. 1; 2, London: Macmillan, 1979.

**Elster, John,** *Ulyssess and the Sirens*, Cambridge: Cambridge University Press, 1979.

**Fellner, William,** "Distortion of Subjective Probabilities as a Reaction to Uncertainty," *Quarterly Journal of Economics*, November 1961, *75*, 670–90.

**Ferguson, Thomas S.,** *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press, 1967.

**Fishburn, Peter C.,** "Transitive Measurable Utility," Discussion Paper No. 224, Bell Laboratories, 1981.

**Fischer, Stanley,** "Long Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–206.

**Friedman, Milton,** "The Methodology of Positive Economics," in his *Essays in Positive Economics*, 1953; reprinted in W. Breit and H. M. Hochman, eds., *Readings in Microeconomics*, New York: Holt, Rinehart & Winston, 1968, 23–47.

______, *A Program for Monetary Stability*, New York: Fordham University Press, 1969.

______ **and Savage, L. J.,** "The Utility Analysis of Choices Involving Risks," *Journal of Political Economy*, August 1948, *56*, 279–304.

**Gibson, J.,** "The Theory of Affordances," in R. E. Shaw and J. Bransford, eds., *Perceiving, Acting, and Knowing*, Hillsdale: Lawrence Erlbaum Assoc., 1977.

**Gluckman, M.,** *Essays on the Ritual of Social Relations*, Manchester: Manchester University Press, 1962.

**Gould, Stephen J. and Eldredge, Niles,** "Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered," *Paleobiology*, January 1977, *3*, 115–51.

**Graham, Virginia L. and Tulcea, C. Ionescu,** *A Book on Casino Gambling*, New York: Pocket Books, 1978.

**Green, David M. and Swets, John A.,** *Signal Detection Theory and Psychophysics*, New York: Robert Kriegur, 1974.

**Grinder, R.,** "Parental Childrearing Practices, Conscience, and Resistance to Temptation of Sixth-Grade Children," *Child Development*, December 1962, *33*, 802–20.

**Hahn, Frank,** *On the Notion of Equilibrium in Economics*, Cambridge: Cambridge University Press, 1973.

**Hamilton, W. D.,** "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology*, 1964, *7*, 1–17.

**Hayek, F. A.,** *Studies in Philosophy, Politics, and Economics*, Chicago: University of Chicago Press, 1967.

______, *Law, Legislation, and Liberty*, Chicago: University of Chicago Press, 1973.

**Heiner, Ronald A.,** "A Theory of Predictable Behavior: Application to Insurance Behavior Anomolies," Department of Economics, Brigham Young University, February 1982.

**Heisenberg, Werner,** *The Physical Principles of the Quantum Theory*, New York: Dover, 1949.

**Herrnstein, Richard J.**, "On the Law of Effect," *Journal of the Experimental Analysis of Behavior*, November 1970, *13*, 243–66.

______, "Relative and Absolute Strength of Response as a Function of Frequency of Reinforcement," *Journal of the Experimental Analysis of Behavior*, 1961, *4*, 267–72.

______, "Secondary Reinforcement and Rate of Primary Reinforcement," *Journal of the Experimental Analysis of Behavior*, January 1964, *7*, 74–91.

______ **and Heyman, Gene M.**, "Is Matching Compatible with Maximization in Concurrent Variable Interval, Variable Ratio?," *Journal of the Experimental Analysis of Behavior*, March 1979, *31*, 209–23.

______ **and Vaughan, W.**, "Melioration and Behavioral Allocation," in J. E. R. Staddon, ed., *Limits to Action: The Allocation of Individual Behavior*, New York: Academic Press, 1980, 143–76.

**Hess, W. R.**, *Imprinting: Early Experience and the Developmental Psychobiology of Attachment*, New York: Van Nostrand 1973.

**Hetherington, E. Mavis and Park, Ross D.**, *Child Psychology: A Contemporary Viewpoint*, 2d ed., New York: McGraw-Hill, 1979.

**Hey, John D.**, *Uncertainty in Microeconomics*, New York: New York University Press, 1979.

**Heyman, Gene M. and Luce, R. Duncan**, "Operant Matching is not a Logical Consequence of Maximizing Reinforcement Rate," *Animal Learning and Behavior*, May 1979, *7*, 133–40.

**Hicks, John**, "A Suggestion for Simplifying the Theory of Money," *Economica*, February 1935, *2*, 1–19.

**Hirshleifer, Jack**, "Evolutionary Models in Economics and the Law: Cooperation Versus Conflict Strategies," *Research in Law and Economics*, 1982, *4*, 1–60.

______, "Economics from a Biological Viewpoint," *Journal of Law and Economics*, April 1977, *20*, 1–54.

______, *Price Theory and Applications*, 2d ed., Englewood Cliffs: Prentice-Hall, 1980.

______ **and Riley, John**, "The Analytics of Uncertainty and Information," *Journal of Economic Literature*, December 1979, *17*, 1375–421.

**Hoffman, Eric and Spitzer, Mathew**, "The Coase Theorem: Some Experimental Tests," *Journal of Law and Economics*, April 1982, *25*, 73–98.

**Hogarth, R. M.**, "Cognitive Processes and the Assessment of Subjective Probability Distributions," *Journal of the American Statistical Association*, June 1975, *70*, 271–94.

**Holmes, Warren G. and Sherman, Paul W.**, "Kin Selection in Animals," *American Scientist*, January-February 1983, *7*, 46–56.

**Howard, N.**, *Paradoxes of Rationality*, Cambridge: MIT Press, 1971.

**Jerison, Harry**, *Evolution of the Brain and Intelligence*, New York: Academic Press, 1973.

**Jones, Robert**, "On the Origin and Development of Media of Exchange," *Journal of Political Economy*, August 1976, *84*, 757–76.

**Judson, Horace F.**, *The Eighth Day of Creation*, New York: Simon & Schuster, 1979.

**Kahneman, Daniel and Tversky, Amos**, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

______ and ______, "The Framing of Decisions and the Psychology of Choice," *Science Magazine*, January 30, 1981, *211*, 453–58.

______ and ______, "The Psychology of Preferences," *Scientific American*, January 1982, *246*, 160–73.

**Kawasaki, Seiichi, McMillan, John and Zimmerman, Klaus F.**, "Disequilibrium Dynamics: An Empirical Study," *American Economic Review*, December 1982, *72*, 992–1004.

**Keeney, Ralph L. and Raiffa, Howard**, *Decisions with Multiple Objectives: Preference and Value Tradeoffs*, New York: Wiley, 1976.

**Keeton, William**, *Biological Science*, 3d ed., New York: Norton, 1980.

**Kirzner, Israel M.**, *Competition and Entrepreneurship*, Chicago: University of Chicago Press, 1973.

**Klein, Benjamin, Crawford, Robert and Alchian, Armen**, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

______ **and Leffler, Keith**, "The Role of Market Performance in Assuring Contractual Performance," *Journal of Political*

*Economy*, October 1981, *89*, 810–34.

**Knight, Frank,** *Risk, Uncertainty and Profit*, Boston: Houghton Mifflin, 1921.

**Kohlberg, Lawrence,** "Stage and Sequence: The Cognitive-Developmental Approach to Socialization," in D. A. Goshn, ed., *Handbook of Socialization Theory and Research*, New York: Rand McNally, 1969, 347–480.

______, "The Development of Children's Orientation Toward Moral Order: Sequence in the Development of Moral Thought," *Vita Humana*, January 1963, *3*, 11–33.

______, "The Domain and Development of Moral Judgment: A Theory and Method of Assessment," in his et al., eds., *Assessing Moral Judgment States: A Manual*, New York: Humanities Press, 1976, 14–45.

______, "Cognitive Stages and Preschool Education," *Human Development*, January 1966, *9*, 5–17.

______, "Justice as Reversibility," in P. Laslett and J. Fishkin, eds., *Philosophy, Politics, and Society*, New Haven: Yale University Press, 1979.

**Konner, M.,** *The Tangled Wing: Biological Constraints on the Human Spirit*, New York: Basic Books, 1982.

**Kosniowski, Czes,** *Conquer that Cube*, Cambridge: Cambridge University Press, 1981.

**Kreps, David et al.,** "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory*, August 1982, *27*, 245–52.

**Kuhn, Thomas,** *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.

**Kunreuther, Howard et al.,** *Disaster Insurance Protection*, New York: Wiley, 1978.

**Kydland, Finn E. and Prescott, Edward C.,** "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

**Lakatos, Imer and Musgrave, Alice,** *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1970.

**Langlois, Richard,** "Internal Organization in a Dynamic Context: Some Theoretical Considerations," Economic Research Report No. 83–04, C. V. Starr Center for Applied Economics, New York University, January 1983.

**Leamer, Edward,** "'Explaining Your Results' As Access Biased Memory," *Journal of the American Statistical Association*, March 1975, *70*, 88–93.

______, *Specification Searches: Ad-Hoc Inference With Nonexperimental Data*, New York: Wiley, 1978.

**Leblebici, Huseyin and Salanik, Gerald R.,** "Effects of Environmental Uncertainty on Information and Decision Processes in Banks," *Administrative Science Quarterly*, December 1981, *26*, 578–96.

______ **and** ______, "Stability in Interorganizational Exchange: Rulemaking Processes of the Chicago Board of Trade," *Administrative Science Quarterly*, June 1982, *27*, 227–42.

**Leibenstein, Harvey,** "Allocative Efficiency vs. X-Efficiency," *American Economic Review*, June 1966, *56*, 392–415.

**Leijonhufvud, Axel,** *On Keynesian Economics and the Economics of Keynes*, New York: Oxford University Press, 1968.

______, *Information and Coordination*, New York: Oxford University Press, 1981.

______, "Life Among the Econ," *Western Economic Journal*, September 1973, *11*, 327–37.

**Lessa, William and Vogt, Evon,** *Reader in Comparative Religion: An Anthropological Approach*, New York: Harper & Row, 1979.

**Levins, Richard,** *Evolution in Changing Environments*, Princeton: Princeton University Press, 1968.

**Lippman, Stephen and McCall, John,** "The Economics of Job Search: A Survey," *Economic Inquiry*, June 1979, *14*, 155–89.

**Loomes, Graham and Sugden, Robert** "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," Department of Economics Working Paper, University of Newcastle-upon-Tyne, 1982.

______ **and** ______, "A Rationale for Preference Reversal," *American Economic Review*, June 1983, *73*, 428–32.

**Lorenz, Konrad,** *The Foundations of Ethology*, New York: Springer-Verlag, 1981.

**Lucas, Robert E., Jr.,** "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, December 1975, *83*, 1113–44.

______, "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April

1972, *4*, 103–24.

______, "Rules, Discretion, and the Role of the Economics Advisor," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980, 199–210.

______, *Studies in Business Cycle Theory*, Cambridge: MIT Press, 1981.

**Lumsden, Charles J. and Wilson, Edward O.**, *Genes, Mind, & Culture: The Coevolutionary Process*, Cambridge: Harvard University Press, 1981.

**McFarland, David**, *The Oxford Companion to Animal Behavior*, New York: Oxford University Press, 1982.

**Machina, Mark**, "Expected Utility Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 277–323.

**March, James G.**, "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics*, Autumn 1978, *9*, 587–608.

**Masterton, R., Hodos, William and Jerison, Harry**, *Evolution, Brain, and Behavior: Persistent Problems*, New York: Wiley, 1976.

**Maynard Smith, John**, "Group Selection and Kin Selection," *Nature*, March 14, 1964, *201*, 1145–47.

**Mayr, Ernst**, *Populations, Species, and Evolution*, Cambridge: Belknap, 1970.

______ **and Provine, W.**, *The Evolutionary Synthesis: Perspectives on the Unification of Biology*, Cambridge: Harvard University Press, 1980.

**Mazur, John E.**, "Optimization Theory Fails to Predict Performance of Pigeons in a Two-Response Situation," *Science*, September 1981, *214*, 823–5.

**Menger, Carl**, *Principles of Economics*, (1871), trans. by James Dingwall and Bert F. Hozelitz, eds., New York: New York University Press, 1981.

______, *Problems In Economics and Sociology*, (1883), trans. by F. J. Nock, Urbana: University of Illinois Press, 1963.

**Michaels, C. F. and Carello, C.**, *Direct Perception*, Englewood Cliffs: Prentice-Hall, 1981.

**Miller, G. T., Galanter, E. and Pribram, K. H.**, *Plans and the Structure of Behavior*, New York: Holt, 1960.

**Miller, James G.**, *Living Systems*, New York: McGraw-Hill, 1978.

**Mondon, Y.**, "What Makes the Toyota Production System Really Tick," *Industrial Engineering Magazine*, January 1981, *17*, 37–46.

**Monod, Jacques**, *Chance and Necessity*, New York: Random House (Vintage Books), 1972.

**Montgomery, Viscount**, *A History of Warfare*, Cleveland: World Publishing, 1968.

**Montross, L.**, *War Through the Ages*, New York: Harper & Brothers, 1960.

**Muth, Richard**, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

**Nelson, Richard**, "Research on Productivity Growth and Productivity Differences: Dead Ends and New Departures," *Journal of Economic Literature*, September 1981, *19*, 1029–64.

______ **and Winter, Sidney**, *An Evolutionary Theory of Economic Capabilities and Behavior*, Cambridge: Harvard University Press, 1982.

______ **and** ______, "Neoclassical Versus Evolutionary Theories of Economic Growth," *Economic Journal*, December 1974, *84*, 886–905.

**Newell, A. and Simon, Herbert**, *Human Problem Solving*, Englewood Cliffs: Prentice-Hall, 1972.

**Nourse, James G.**, *The Simple Solution to Rubik's Cube*, New York: Bantam Books, 1980.

**Okun, Arthur M.**, "Inflation: Its Mechanics and Welfare Cost," *Brookings Papers on Economic Activity*, 2: 1975, 351–401.

______, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

**Ouchi, William**, *Theory Z*, New York: Avon Publishers, 1981.

**Pauli, Wolfgang**, *Theory of Relativity*, New York: Dover, 1981.

**Phelps, Edmond**, "Okun's Micro-Macro System: A Review Article," *Journal of Economic Literature*, September 1981, *19*, 1065–73.

**Piaget, J.**, *The Psychology of Intelligence*, London: Routledge & Kegan Paul, 1947.

______, *The Origins of Intelligence*, New York: International Universities Press, 1952.

**Pianka, Eric R.,** *Evolutionary Ecology*, 2d ed., New York: Harper & Row, 1978.

**Pilbean, D.,** *The Ascent of Man: An Introduction to Human Evolution*, New York: Macmillan, 1972.

**Popper, Karl,** *Conjectures and Refutations: The Growth of Scientific Knowledge*, 3d ed., rev., London: Routledge & Kegan Paul, 1969.

______, *The Logic of Scientific Discovery*, New York: Basic Books, 1959.

**Posner, Richard,** *Economic Analysis of Law*, 2d ed., Boston: Little-Brown, 1977.

______, "A Theory of Primative Society, With Special Reference to Primative Law," *Journal of Law and Economics*, April 1980, *23*, 1–54.

**Prelec, D.,** "Matching, Maximizing, and the Hyperbolic Reinforcement Feedback Function," *Psychological Review*, March 1982, *89*, 189–230.

**Premack, David,** *The Mind of an Ape*, New York: Norton, 1983.

**Priest, George L.,** "The Common Law Process and the Selection of Efficient Rules," *Journal of Legal Studies*, January 1977, *6*, 65–83.

**Rachlin, Howard, Kagel, John H. and Battalio, R. C.,** "Substitutability in Time Allocation," *Psychological Review*, July 1980, *87*, 355–74.

**Rawls, John,** *A Theory of Justice*, Cambridge: Harvard University Press, 1971.

**Revere, Lawrence,** *Playing Blackjack as a Business*, Secaucus: Lyle Stuart, Inc., 1969.

**Roughgarden, Jeffrey D.,** "Reasons and Rules in Choice: A Framework for Analysis," Department of Engineering Economics Systems, Stanford University, December 1982.

**Rowan, B.,** "Organizational Structure and the Institutional Environment: The Case of Public Schools," *Administrative Science Quarterly*, June 1982, *27*, 259–79.

**Rozin, P.,** "The Evolution of Intelligence and Access to the Cognative Unconscious," in J. A. Spague and A. N. Epstein, eds., *Progress in Psychobiology and Physiological Psychology*, Vol. 6, New York: Academic Press, 1976, 245–80.

**Sagan, Carl,** *The Dragons of Eden*, New York: Random House, 1977.

**Samuelson, Paul A.,** *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

**Sargent, Thomas J.,** *Macroeconomic Theory*, New York: Academic Press, 1979.

______, (1976a) "A Classical Macroeconometric Model for the United States," *Journal of Political Economy*, March-April 1976, *84*, 207–37.

______, (1976b) "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *Journal of Political Economy*, May-June 1976, *84*, 631–40.

______ **and Wallace, Neil,** "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, *83*, 241–54.

**Savage, L. J.,** *The Foundations of Statistics*, New York: Wiley, 1954.

**Schelling, Thomas C.,** *Micromotives and Macrobehavior*, New York: W. W. Norton, 1978.

**Schotter, Andrew,** *The Economic Theory of Social Institutions*, New York: Cambridge University Press, 1981.

**Schumpeter, Joseph,** *Capitalism, Socialism, and Democracy*, New York: Harper & Brothers, 1942; Harper Colophon Edition, 1976.

**Seligman, Martin E. P.,** "On the Generality of the Laws of Learning," *Psychological Review*, September 1970, *77*, 406–18.

**Shackle, G. L. S.,** *Decision, Order, and Time in Human Affairs*, 2d ed., Cambridge: Cambridge University Press, 1969.

______, *Epistemics and Economics: A Critique of Economic Doctrines*, Cambridge: Cambridge University Press, 1972.

**Simon, A. and Sikossy, L.,** *Representation and Meaning: Experiments With Information Processing Systems*, Englewood Cliffs: Prentice-Hall, 1972.

**Simon, Herbert,** "A Behavioral Theory of Rational Choice," *Quarterly Journal of Economics*, February 1955, *69*, 99–118.

______, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, 2d ed., New York: Macmillan, 1959.

______, *The Sciences of the Artificial*, Cambridge: MIT Press, 1969.

______, "From Substantive to Procedural Rationality," in S. Latsis, ed., *Method and Appraisal in Economics*, Cambridge: Cambridge University Press, 1976, 129–48.

______, *The New Science of Management Decision*, Englewood Cliffs: Prentice-Hall, 1977.

______, "On How to Decide What to Do," *Bell Journal of Economics*, Autumn 1978, *9*, 494–507.

______, (1979a) *Models of Thought*, New Haven: Yale University Press, 1979.

______, (1979b) "Rational Decision Making in Business Organizations," *American Economic Review*, September 1979, *69*, 493–513.

**Singmaster, David** *Notes on Rubik's Magic Cube*, Hillside: Enslow, 1979.

______ **and Frey, Alexander H., Jr.,** *Handbook of Cubic Math*, Hillside: Enslow, 1982.

**Slovic, Paul and Tversky, Amos,** "Who Accepts Savage's Axiom," *Behavioral Science*, November 1974, *19*, 368–73.

______ **and Lichtenstein, Sarah,** "Preference Reversals: A Broader Perspective," *American Economic Review*, September 1983, *73*, 596–605.

**Smith, Adam,** *The Theory of Moral Sentiments*, (1759), New Rochelle: Arlington House, 1969.

**Smith, Vernon L.,** "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, *72*, 923–55.

**Sowell, Thomas,** *Knowledge and Decisions*, New York: Basic Books, 1980.

**Stanley, Steven M.,** *Macroevolution: Pattern and Process*, San Francisco: W. H. Freeman, 1979.

**Staw, Barry, Sanderlands, Lana and Dutton, Jane,** "Threat Rigidity Effects in Organizational Behavior: A Multilevel Analysis," *Administrative Science Quarterly*, December 1981, *26*, 501–24.

**Stigler, George, J.,** "The Economics of Information," *Journal of Political Economy*, June 1961, *69*, 213–25.

**Strotz, R. H.,** "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, November 1955, *23*, 165–80.

**Sugimori, Y., Kusunoki, K. and Cho, S.,** "Toyota Production System and Kanban System: Materialization of 'Just in Time' Production and 'Respect for Human' System," *International Journal of Production Research*, December 1977, *15*, 553–64.

**Sussman, Hector J.,** "Catastrophe Theory," *Synthese*, August 1975, *31*, 229–70.

**Taylor, D.,** *Mastering Rubik's Cube*, New York: Holt, Rinehart & Winston, 1980.

**Telser, Lester,** "Why Are There Organized Futures Markets?," *Journal of Law and Economics*, April 1981, *24*, 1–22.

**Thòm, Rene,** *Structural Stability and Morphogenesis*, Reading: W. A. Benjamin, 1975.

**Thorpe, Edward, O.,** *Beat the Dealer: A Winning Strategy for the Game of Twenty-One*, New York: Vintage Books, 1962.

**Tinbergen, Nino,** *The Study of Instinct*, London: Oxford University Press, 1951.

**Toates, Fredric,** *Animal Behavior: A Systems Approach*, New York: Wiley, 1980.

**Tobin, James,** "Are New Classical Models Plausible Enough to Guide Policy?," *Journal of Money, Credit and Banking*, November 1980, *12*, 788–99.

**Trivers, Robert L.,** "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, March 1971, *46*, 35–58.

**Tversky, Amos,** "Intransitivity of Preferences," *Psychological Review*, January 1969, *76*, 31–48.

**Ullman-Margalitt, Edna,** *The Emergence of Norms*, New York: Oxford University Press, 1978.

**Varian, Hal R.,** "Catastrophe Theory and the Business Cycle," *Economic Inquiry*, January 1979, *17*, 14–28.

**Vaughan, William,** "Melioration, Matching, and Maximization," *Journal of the Experimental Analysis of Behavior*, September 1981, *36*, 141–49.

**von Neumann, John and Morgenstern, Oskar,** *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1944.

**Watson, James, D.,** *Molecular Biology of the Gene*, 3d ed., Menlo Park: W. A. Benjamin, 1976.

**Weber, R.,** "The Allais Paradox, Dutch Auctions, and Alpha-Utility Theory," J. L. Kellogg Graduate School of Management Working Paper, Northwestern University, 1982.

**Weiner, Norbert,** *Cybernetics*, Cambridge: MIT Press, 1948.

**Weintraub, E. Roy,** "The Microfoundations of Macroeconomics: A Critical Survey," *Journal of Economic Literature*, March 1977, *15*, 1–23.

**Wendt, Dirk,** "Some Criticism of Stochastic Models Generally Used in Decision Making Experiments," *Theory and Decision*, May 1975, *6*, 197–212.

**White, Andrew D.,** *Fiat Money Inflation in France*, New York: D. Appleton-Century, 1933; Los Angeles: Pamphleteers, Inc., 1945.

**Wilkins, Maurice,** "Molecular Structure of Deoxypentos Nucleic Acids," *Nature*, April 1953, *171*, 738–40.

**Williamson, Oliver E.,** *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: The Free Press, 1975.

______, "Transactions-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, *22*, 233–61.

______, "The Modern Corporation: Origins, Evolution, Attributes," *Journal of Economic Literature*, December 1981, *19*, 1537–68.

______, "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, *73*, 519–40.

**Wilson, Edward O.,** *On Human Nature*, Cambridge: Harvard, 1978.

______, *Sociobiology: The New Synthesis*, Cambridge: Belknap, 1975.

**Winter, Sidney G.,** "Economic 'Natural Selection' and the Theory of the Firm," *Yale Economic Essays*, May 1964, *4*, 225–72.

______, "Satisficing, Selection, and the Innovating Remnant," *Quarterly Journal of Economics*, May 1971, *85*, 237–62.

______, "Optimization and Evolution," in R. H. Day and R. Groves, eds., *Adaptive Economic Models*, New York: Academic Press, 1975.

**Winterfeldt, Detlof,** "Additivity and Expected Utility in Risky Multi-Attribute Preferences," *Journal of Mathematical Psychology*, February 1980, *21*, 66–82.

**Zeeman, E. C.,** *Catastrophy Theory*, New York: Addison-Wesley, 1977.

# Preference Reversals: A Broader Perspective

*By* PAUL SLOVIC AND SARAH LICHTENSTEIN*

Two papers recently published in this *Review*, the first by Werner Pommerehne, Freidrich Schneider, and Peter Zweifel (1982) and the second by Robert Reilly (1982), reexamined the preference reversal phenomenon. Preference reversals occur when individuals are presented with two gambles, one featuring a high probability of winning a modest sum of money (the *P* bet), the other featuring a low probability of winning a large amount of money (the $ bet). The typical finding is that people often choose the *P* bet but assign a larger monetary value to the $ bet. This behavior is of interest because it violates almost all theories of preference, including expected utility theory.

The studies by Pommerehne et al. and Reilly were based on an earlier paper appearing in this *Review* by David Grether and Charles Plott (1979). All three of these investigations have followed the same general design, motivated by a healthy skepticism of the phenomenon and a belief that, examined under proper conditions, it might disappear. Thus Grether and Plott took great pains to correct what they saw as deficiencies in the original psychological experiments by ourselves (1971, 1973) and Harold Lindman (1971). Specifically, Grether and Plott used two monetary incentive systems to heighten motivation, substituted a different probability device for deciding the outcomes of the bets, controlled for income and order effects, and tested for indifference and the influence of strategic or bargaining effects. To their surprise, preference reversals remained much in evidence, despite their careful attempts to create conditions that would minimize or eliminate them.

Pommerehne et al., not satisfied with the stringency of Grether and Plott's controls, attempted to increase motivation by raising the face value of the payoffs and creating differences in expected value between the *P* and $ bets in a pair. They, too, found a substantial proportion of reversals, leading them to conclude: "Even when the subjects are exposed to strong incentives for making motivated, rational decisions, the phenomenon of preference reversal does not vanish" (p. 573).

Reilly was also skeptical of the adequacy of Grether and Plott's controls. To maximize subjects' understanding of the task, he conducted his study within small groups where questions could readily be asked of the experimenter. The money at risk was placed on a desk in front of the subject and the size of potential losses in the gambles was increased to enhance motivation. Finally, some subjects were shown the expected values for all gambles and were given a description of the expected-value concept. Although the rate of preference reversals was somewhat lower than that observed by Grether and Plott, the phenomenon persisted to a substantial extent. Reilly conceded that these results provided "further confirmation of preference reversal as a persistent behavioral phenomenon in situations where economic theory is generally applied" (p. 582). Nevertheless, he maintained the hope that further strengthening of monetary incentives and provision of additional information to the subjects would make this troublesome phenomenon disappear, thus salvaging preference theory:

> Should sufficiently large reductions be achievable, we might consider adopting the premise that individuals are likely to be consistent in making decisions that *matter* to them when the principle characteristics of the alternatives are sufficiently comprehended. Applied to such cases, standard preference theory would then require little modification. [p. 582]

---

*Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, Oregon 97401. The work was supported by the Office of Naval Research under Contract N00014-82-C-0643 to Perceptronics, Inc. We thank Don MacGregor, Amos Tversky, and two anonymous reviewers for comments on an earlier draft.

As researchers who have studied preference reversals and related problems of rational choice for quite some time, we have several concerns about the direction this research seems to be taking. Certainly a phenomenon such as preference reversals should be subjected to rigorous tests such as those administered by Grether and Plott, Pommerehne et al., and Reilly. These studies have been valuable in demonstrating the robustness of the effect. However, there is a substantial body of research on preference reversals within the psychological literature that is being neglected here. Moreover, reversals can be seen not as an isolated phenomenon, but as one of a broad class of findings that demonstrate violations of preference models due to the strong dependence of choice and preference upon information processing considerations. In this paper we shall describe relevant psychological work in order to broaden the perspective on preference reversals.

## I. History

Readers of the papers by Pommerehne et al. and Reilly would hardly know there was considerable scrutiny of preference reversals prior to the publication by Grether and Plott. In fact, a number of studies preceded Grether and Plott, most of which employed multiple experiments and conditions designed to test the robustness of the effect. Additional studies have appeared subsequently. Each of these studies has observed substantial frequencies of reversals.

The first study designed to elicit reversals was our 1971 article. The impetus for this study was our observation in our earlier 1968 article that choices among pairs of gambles appeared to be influenced primarily by probabilities of winning and losing, whereas buying and selling prices were primarily determined by the dollar amounts that could be won or lost. When subjects found a bet attractive, their prices correlated predominantly with the amount to win; when they disliked a bet, their prices correlated primarily with the amount that could be lost. This pattern of correlations was explained as the result of a starting point (anchoring) and adjustment procedure used when setting prices. Subjects setting a price on an attractive gamble appeared to start with the amount to win and adjust it downward to take into account the probability of winning and losing, and the amount that could be lost. The adjustment process was relatively imprecise, leaving the price response greatly influenced by the starting point payoff. Choices, on the other hand, appeared to be governed by different rules.

In our 1971 article, we argued that, if the information in a gamble is processed differently when making choices and setting prices, it should be possible to construct pairs of gambles such that people would choose one member of the pair but set a higher price on the other. We proceeded to construct a small set of pairs that clearly demonstrated this predicted effect.[1] Following this, a second study was conducted to examine the strength of the reversal effect as a function of the characteristics of the bet pairs. Forty-nine pairs of bets were constructed, all constrained by the requirement that the *P* bet had a high probability of winning a modest amount and the $ bet had a low to moderate probability of winning a large amount. Despite these constraints, the pairs differed significantly in the degree to which they elicited predictable reversals. The ideal bet pair for observing reversals had a larger $ bet loss than a *P* bet loss (facilitating choice of the *P* bet) and a large $ bet win relative to the *P* bet win (facilitating a higher price for the $ bet). For example, the bet with the most predicted reversals was: *P* bet, 9/12 to win $1.20 and 3/12 to lose $.10; $ bet, 3/12 to win $9.20 and 9/12 to lose $2.00. We concluded this initial study by noting that reversals were of interest not only because they violated theories of rational choice, but because of the insight they revealed about the nature of judgment and decision processes.

Our 1968 article also noted that the close dependence of pricing responses on a gamble's payoffs could explain a finding that had puzzled Lindman (1965) in his doctoral dissertation. Lindman's subjects gave selling prices for gambles and also made paired-comparison choices among triplets of these gambles. He noted that the prices were ordered almost perfectly according to the payoffs, whereas the orderings derived from choices were not. Lindman (1971) subsequently performed five studies designed to determine whether this sort of inconsistency would be influenced by the number of gambles within the choice set, the possibility of comparing gambles directly when deciding upon selling prices, variations in the way that probabilities were displayed, and variations in the amount of prior practice or experience. Although the experience factor had some effect, the general results across conditions were in close agreement with our own.

---

[1] Contrary to the explanation by Reilly, the act of choosing the *P* bet but setting a higher price on the $ bet is not called a *predicted* reversal simply because "In all experiments reversal of *P* bets has been more frequent than reversal of $ bets..." (Reilly, 1982, p. 577, fn. 2).

Problems of motivation and understandability were of concern right from the beginning of these studies. Experiment III of our original paper (1971) allowed college student subjects to win up to $8, a significant amount for an hour's work in 1969. Each subject was run individually, with lengthy and careful instructions. Prices and choices were obtained three times for each pair of bets. The third time, subjects were reminded what their earlier answers had been and were asked to make a careful, final response. The bets were actually played and subjects were paid as a function of their winnings. Results for these carefully trained and financially motivated subjects showed a substantial proportion of predicted reversals. Recognizing the importance of motivation and the need to test nonstudent subjects, we went to considerable effort to replicate the initial studies on the floor of a casino in downtown Las Vegas.[2] There the players could set the value of their chips at $.05, $.10, $.25, $1, or $5. No players ever chose $1 or $5, but even for the $.10 chips, a typical $ bet offered either a win or a loss of $8 on a single play. One new feature of the design was the addition of gambles having negative expected values. The experiment attracted 44 players, many of whom were highly educated professionals. Reversals of preference were frequent and widespread across players, even for the negative expected value gambles, for which strategic tendencies to overprice the bets would have worked against the reversal phenomenon.

Robert Hamm (1979) was another researcher who tried hard to make the reversal phenomena disappear—and did not. His extensive study examined the stability of reversals over time, in the face of experience, practice, forced introspection or discussion, and advice to adopt an intuitive or analytic approach to the task. The order of stimulus sets and tasks was carefully counterbalanced. Hamm found that the reversal effect was replicated under all these conditions. Task order had no effect, nor did emphasis on analytic or intuitive processes. Discussion about one's decision strategies actually increased the tendency towards reversals, countering the hypothesis that if people were given greater opportunity to think about their strategies, the preference reversal phenomenon would disappear.

John Mowen and James Gentry (1980) studied preference reversals in a quite different context—that of new product development. Their subjects were undergraduate students of marketing and consumer behavior. They also extended previous research by comparing individual vs. group decisions. The stimuli were hypothetical products, defined according to probability of success and failure, and the projected profits and losses associated with those probabilities. Although the proportion of reversals varied with the characteristics of the pairs, as found in our (1971) study, strong reversal effects were generally observed. Group judgments and decisions were even more prone to reversals than those of individuals. Because group decisions involve discussion of strategies, this result is congruent with the effects of discussion found by Hamm. Mowen and Gentry related the anchoring process thought to determine pricing responses to an anecdote provided by R. A. Kerr (1979) regarding the search for oil in the Baltimore Canyon. Kerr noted that oil companies paid $1.1 billion for the privilege of drilling despite negative reports from oil industry geochemists. He concluded that "Company managers apparently bid more on the basis of how large the possible trapping structures were rather than on the basis of the odds figured by the geochemists" (p. 1071).

---

[2] See our article (1973).

In sum, many of the concerns raised and examined by Grether and Plott have also been investigated in other studies of preference reversals. Our purpose in reviewing these studies is not to deny the importance of the studies by Grether and Plott, Pommerehne et al., and Reilly, but rather to inform those interested in this topic about the larger body of results. In our opinion, the most striking result of these studies is the persistence of preference reversals in the face of determined efforts to minimize or eliminate them.

## II. A Broader View of Preference Reversals

The inconsistency between prices and choices for risky prospects represents but one of a broad set of failings that have been attributed to the theory of rational choice. James March (1978, 1982) has identified five general problems with the theory, one of which is particularly relevant to the present discussion. According to March, the theory presumes two improbably precise guesses about the future. One is a guess about the future consequences of current actions. The other is a guess about future sentiments (i.e., preferences) with respect to those consequences.

March (1978) argued that, partly as a result of behavioral research on human information-processing limitations, the way that the rational theory deals with the first guess has been modified to incorporate principles of what Herbert Simon (1957) termed "bounded rationality." Thus economic theories now place considerable emphasis on notions of search, attention, and information costs. Aspiration levels, incrementalism, and satisficing have been described as sensible in many settings.

In contrast, March observed that although the second guess, about uncertain preferences, has so far had little effect in modifying normative theories, it poses potentially greater difficulties for these theories and their applications. He argued that limited cognitive capacity affects information processing about preferences just as it affects information processing about consequences: "Human beings have unstable, inconsistent, incompletely evoked, and imprecise goals at least in part because human abilities limit preference orderliness" (1978, p. 598).

March draws upon a rich and diverse array of observations to argue that, contrary to normative theory, preferences are neither absolute, stable, consistent, precise or exogenous (unaffected by the choices they control). The case against consistency brings us back to the topic of preference reversals. Inconsistencies between prices and choices were created on the basis of knowledge about different rules for processing the component aspects or dimensions of gambles. Since 1968, when information processing ideas began to be applied to risky choice, we have learned more about how perception and cognition determine preferences. As we have better understood those processes, it has become relatively easy, indeed almost commonplace, to produce new kinds of preference reversals. In many instances, production of reversals has been used to validate hypotheses about information processing in risky choice.

An early demonstration of the link between information processing and reversals was a study by Amos Tversky (1969). He hypothesized that, where the structure of the choice set permitted, it would be simpler and more natural to compare alternatives dimension by dimension than to evaluate the combined worth of each alternative separately (across dimensions) and then compare these overall evaluations. Tversky further hypothesized that small differences (for example, below some threshold of discrimination) would be ignored, even for an important dimension. Tversky tested and confirmed these hypotheses by creating sets of gambles in which this sort of information processing led to systematic, predictable intransitivities. Tversky's gambles contained only two dimensions, probability of winning and amount to win. For his subjects, probability was the dominant dimension, but if the difference between gambles was small, then amount to win controlled the decision. Thus, given the set of gambles *a*, *b*, *c*, *d*, and *e* with probabilities of 7/24, 8/24, 9/24, 10/24, and 11/24 to win $5.00, $4.75, $4.50, $4.25, and $4.00, respectively, *a* tended to be chosen over *b*, *b* over *c*, *c* over *d*, and *d* over *e*, presumably because the difference in payoffs outweighed the slight difference in probabilities within each of these pairs. However, *e* was typically chosen over *a* because of the relatively large difference in probabilities. This general finding has subsequently been replicated and extended by Rob Ranyard (1976) and by Lindman and James Lyons (1978).

The intransitivities observed by Tversky arose from the tendency of subjects to compare gambles on each dimension. If they had made holistic evaluations separately for each gamble and compared these to determine their choices, then the intransitivities would not have occurred. Comparison within dimensions is a natural way to choose among multidimensional objects. However, information is sometimes not available for each dimension, a situation that can lead to reversals. Consider, for example, the task of predicting which of two college students, *A* or *B*, would get the higher grade point average. Two test scores are available for each student, to serve as the basis for prediction. One score, English Skill, is available for both students. The other information is unique —Quantitative Ability for Student *A* and Need for Achievement for Student *B* as shown below (the means and standard deviations of each test are different but are known to the evaluator).

| | Student *A* | Student *B* |
|---|---|---|
| Need for Achievement | – | 30 |
| English Skills | 90 | 131 |
| Quantitative Ability | 602 | – |

Slovic and Douglas MacPhillamy (1974) hypothesized that commonality would cause a dimension to be weighted more heavily in determining a choice, because common information is easier to use. This, in fact, occurred and led to systematic reversals on the above problem: 75 percent of the subjects rating the students individually gave a higher grade point average to Student *A*. However, when these same subjects were asked to make a comparative judgment, they selected Student *B* 60 percent of the time (40 percent of the subjects exhibited reversals). Reversals also occurred, though less frequently, when the means and standard deviations were the same for each test.

A variety of different reversals, providing strong evidence against traditional theories of preference, have come from the work of Daniel Kahneman and Tversky (1979; Tversky and Kahneman, 1981). From their systematic observations of choices among risky alternatives, Kahneman and Tversky have deduced a number of general principles, some of which violate expected utility theory, others of which are incompatible with all existing theories of choice or preference. Kahneman and Tversky distinguished between two phases in the choice process, an early phase of editing and a subsequent phase of evaluation. The editing phase, which they have also referred to as *framing*, consists of a preliminary analysis of the available options, their possible outcomes, and the contingencies or conditional probabilities relating outcomes to acts. One function of the framing process is to organize and reformulate the alternatives so as to simplify the second phase of evaluation and choice. Much as changes in vantage point induce alternative perspectives on a visual scene, the same decision problem can be subject to many alternative frames. Whichever frame is adopted is determined in part by the external formulation of the problem and in part by the standards, habits, and personal predilections of the decision maker.

A key element of framing is the coding of outcomes. Kahneman and Tversky show that, contrary to utility theory, outcomes are typically coded as gains and losses, rather than as final states of wealth. These gains and losses are defined relative to some neutral reference point, typically, but not always, the current asset position of the decision maker. These changes are evaluated according to a value function, $v(x)$, which attaches a subjective worth to each possible outcome of a gamble, and a nonlinear probability weighting function, $\pi(p)$, which expresses the subjective importance attached to the probability of obtaining a particular outcome. The attractiveness of a gamble that offers a chance of $p$ to obtain outcome $x$ and a chance of $q$ to obtain outcome $y$ would be equal to $\pi(p)v(x)+\pi(q)v(y)$. In addition to being defined on gains and losses relative to some psychologically meaningful (neutral) reference point, the value function is steeper for losses than for gains, meaning that a given change in one's status hurts more as a loss than it pleases as a gain. Another important feature is that the function is concave above that reference point and convex below it, meaning, for example, that the subjective difference between gaining (or losing) \$10 and \$20 is greater than the difference between gaining (or losing) \$110 and \$120. Perhaps the most notable feature of the probability weighting function is the great importance attached to outcomes that will be received with certainty. Thus, for example, the prospect of losing \$50 with probability of 1.0 is more than twice as aversive as the prospect of losing the same amount with probability .5.

The way a problem is framed determines both the reference point (the zero point) of the value function and the probabilities that are evaluated. If $\pi$ and $v$ were linear functions, preferences among options would be independent of the framing of acts, outcomes, or contingencies. Because of the characteristic nonlinearities of $\pi$ and $v$, however, normatively inconsequential changes in the frames significantly affect preferences. This is illustrated by the following pair of problems, given to separate groups of respondents.

*Problem* 1. Imagine that the United States is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the consequences of the programs are as follows: If Program *A* is adopted, 200 people will be saved. If Program *B* is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. Which of the two programs would you favor?

*Problem* 2. (Same cover story as Problem 1.) If Program *C* is adopted, 400 people will die. If Program *D* is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. Which of the two programs would you favor?

Although the two problems are formally identical, the preferences tend to be quite different. In a study of college students, 72 percent of the respondents chose Program *A* over Program *B* and 78 percent chose Program *D* over Program *C*. This difference can be traced to the different frames implied by the two problems. The "save lives" wording of the first problem implies that the value function's reference point is the loss of 600 lives, while the "people will die" wording of problem 2 suggests that the reference point is at no lives lost. Thus problem 1 falls in the concave gain region of the value function while problem 2 is in the convex loss region. Another study, surveying physicians and patients regarding choice of radiation vs. surgical treatments for lung cancer, produced different decisions when relevant statistics were changed from probabilities of surviving for various lengths of time after treatment to probabilities of not surviving (Barbara McNeil et al., 1982).

Another example of framing effects has been presented by Kahneman and Tversky (1982).

*Problem* 1. Imagine that, in addition to whatever else you have, you have been given \$200. You are now asked to choose between (A) a sure gain of \$50 and (B) a 25 percent chance of winning \$200 and a 75 percent chance of winning nothing.

*Problem* 2. Imagine that, in addition to whatever you have, you have been given a cash gift of \$400. You are now asked to choose between (C) a sure loss of \$150, and (D) a 75 percent chance of losing \$200 and a 25 percent chance of losing nothing.

Most people choose A over B and D over C. Yet, the options presented in the two problems are identical. There is no valid reason to prefer the gamble in one version and the sure outcome in the other. Choosing the sure gain in the first problem yields a total gain of \$200 plus \$50, or \$250. Choosing the sure loss in the second version yields the same result through the deduction of \$150 from \$400. The choice of the gamble in either problem yields a 75 percent chance of winning \$200 and a 25 percent chance of winning $400. If the respondents to these problems took a comprehensive view of the consequences, as is assumed by theories of rational decision, they would combine the bonus with the available options and evaluate the composite outcome. Instead they ignore the bonus and evaluate the first problem as a choice between gains and the second as a choice between losses. The reversal of preferences is induced by reframing the problem.

We have used the framing and reference point notions to explain the finding that the certain loss of a stated amount of money (for example, $50) was much more attractive when described as an insurance premium (to safeguard against a .25 chance of losing $200) than when it was described as an alternative to playing that same gamble (see our article with Baruch Fischhoff, 1982a; see also Paul Schoemaker and Howard Kunreuther, 1979, and John Hershey and Schoemaker, 1980, for similar results).

### III. Where Next?

We have presented a sample of the sorts of preference reversals that have formed our understanding of choice processes or have been created from that understanding. Those who are concerned about the possible economic implications of these phenomena have several paths to consider. One is to continue to subject these studies to the sorts of scrutiny that Grether and Plott and others have applied to the inconsistency between prices and choices. Despite the claims by Tversky and Kahneman (1981) that the effects they described are large and systematic, associated with losses of human life as well as monetary outcomes, not restricted to hypothetical questions, nor eliminated by monetary incentives, this line of research is young and there is certainly a need to test the limits and robustness of its findings.

A second path is to modify utility theory in order to accommodate as many of the behavioral anomalies as possible without abandoning the theory altogether. This has been a popular direction in recent years. A number of theorists have proposed weakening or eliminating the substitution axiom in order to accommodate the Allais paradox (Maurice Allais, 1953) and certain other violations of the traditional model (see S. H. Chew and Kenneth MacCrimmon, 1979; Peter Fishburn, 1981; Robert Weber, 1982; Hector Munera and Richard de Neufville, 1982; and Mark Machina, 1982). However, none of these revamped models can explain the framing effects described by Tversky and Kahneman (1981) or the preference reversals among *P* bets and $ bets. Indeed, Machina acknowledged that, "to the extent that preference reversals are found to be systematic and pervasive, the behavioral model presented here must either be generalized or replaced" (p. 308).

A third path to follow, and one that we would advocate, is to accept the reality of preference reversals and related information-processing phenomena, and to explore their implications for important social and economic behaviors. We have begun to do this with regard to problems of societal risk management and programs for informing the public about risk (see our study, with Fischhoff, 1982b). Similarly, March (1978, 1982), whose critique went far beyond information processing to encompass complex strategic and social motivations, has urged that a conception of preference that respects the "intelligence of ambiguity" be incorporated into what he calls "the engineering of choice." He identified a number of conceptual problems that need to be addressed by choice theorists and optimization problems that need to be considered by choice engineers.

In a narrower but nonetheless important vein, Hershey, Kunreuther, and Schoemaker (1982) have demonstrated biases in utility functions caused by information processing effects. They showed that methods for assessing utilities, varying in normatively inconsequential ways, produced very different utility functions, posing both practical and theoretical problems for those concerned with assessing people's risk preferences. Donald Wehrung, MacCrimmon, and K. Brothers (1980) obtained similar inconsistencies with business executives, leading them to question the use of utility theory as a management tool. A more general analysis of the difficulties of assessing preferences has been presented by Fischhoff and ourselves (1980).

Fischhoff et al. argue that the strong effects of framing and information-processing considerations make elicitation methods major forces in shaping the expression of one's personal values.

Robin Gregory (1982) investigated a number of different approaches for estimating the value of nonmarket goods such as air and water quality, protection of threatened environments and species, and access to uninhabited views. He examined two measures of economic value; one based on an individual's willingness to pay to obtain or retain a good and the other based on the amount of compensation demanded if it is relinquished. He found that both methods were subject to sizable framing and information-processing effects.

Richard Thaler (1980) has drawn upon the reference point and framing notions of Kahneman and Tversky to explain a number of "economic illusions" that cause consumer behavior to deviate from the predictions of normative models. Included in his analysis were the overweighting of out-of-pocket costs relative to opportunity costs (foregone gains), the failure to ignore sunk costs, and the effects of psychic regret on such diverse areas as health care delivery decisions and vacation planning. Thomas Russell and Thaler (1982) argued that departures from rationality due to information-processing effects are unlikely to disappear in competitive markets. Kenneth Arrow (1982) underscored this argument by pointing out a number of failures of the rational model in insurance, securities, and futures markets that he feels are directly interpretable in terms of effects such as those linked to preference reversals and framing.

## IV. Conclusion

This review has attempted to show how preference reversals fit into a larger picture of information-processing effects that, as a whole, pose a collective challenge to preference theories far exceeding that from reversals alone. These effects seem unlikely to disappear, even under rigorous scrutiny. Moreover, anything less than a radical modification of traditional theories is unlikely to accommodate these phenomena. We urge economists not to resist these developments but, instead, to examine them for insights into the ways that decisions are made and the ways that the practice of decision making can be improved.

## REFERENCES

**Allais, Maurice,** "Le Comportement de l'Homme Rationnel devant le Risque: Critiques des Postulats et Axiomes de l'Ecole Americaine," *Econometrica*, October 1953, *21*, 503–46.

**Arrow, Kenneth J.,** "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, *20*, 1–9.

**Chew, S. H. and MacCrimmon, Kenneth,** "Alpha Utility Theory, Lottery, Composition, and the Allais Paradox," Working Paper No. 686, University of British Columbia, 1979.

**Fischhoff, Baruch, Slovic, Paul and Lichtenstein, Sarah,** "Knowing What You Want: Measuring Labile Values," in Tom Wallsten, ed., *Cognitive Processes in Choice and Decision Behavior*, Hillsdale: Erlbaum, 1980.

**Fishburn, Peter C.,** "Transitive Measurable Utility," Economic Discussion Paper No. 224, Bell Laboratories, Murray Hill, 1981.

**Gregory, Robin S.,** "Valuing Non-Market Goods: An Analysis of Alternative Approaches," unpublished doctoral dissertation, University of British Columbia, June 1982.

**Grether, David M. and Plott, Charles R.,** "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

**Hamm, Robert MacGowan,** "The Conditions of Occurrence of the Preference Reversal Phenomenon," unpublished doctoral dissertation, Harvard University, 1979.

**Hershey, John C., Kunreuther, Howard and Schoemaker, Paul J. H.,** "Bias in Assessment Procedures for Utility Functions," *Management Science*, August 1982, *28*, 936–54.

______ **and Schoemaker, Paul J. H.,** "Risk Taking and Problem Context in the Domain of Losses: An Expected Utility Analysis," *Journal of Risk and Insurance*, March 1980, *47*, 111–32.

**Kahneman, Daniel and Tversky, Amos,** "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

______ **and** ______, "The Psychology of Preferences," *Scientific American*, January 1982, *246*, 160–73.

**Kerr, R. A.,** "Petroleum Exploration: Discouragement about the Atlantic Outer Continental Shelf Deepens," *Science*, June 8, 1979, *204*, 1069–72.

**Lichtenstein, Sarah and Slovic, Paul,** "Reversals of Preference Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, January 1971, *89*, 46–55.

______ **and** ______, "Response-Induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas," *Journal of Experimental Psychology*, November 1973, *101*, 16–20.

**Lindman, Harold R.,** "Simultaneous Measurement of Utilities and Subjective Probabilities," unpublished doctoral dissertation, University of Michigan, 1965.

______, "Inconsistent Preferences Among Gambles," *Journal of Experimental Psychology*, May 1971, *89*, 390–97.

______ **and Lyons, James,** "Stimulus Complexity and Choice Inconsistency Among Gambles," *Organizational Behavior and Human Performance*, April 1978, *21*, 146–59.

**McNeil, Barbara J. et al.,** "On the Elicitation of Preferences for Alternative Therapies," *New England Journal of Medicine*, May 1982, *306*, 1259–62.

**Machina, Mark, J.,** "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 277–323.

**March, James G.,** "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics*, Autumn 1978, *9*, 587–608.

______, "Theories of Choice and the Making of Decisions," presented at the Annual Meeting of the American Association for the Advancement of Science, Washington, D.C., January 5, 1982.

**Mowen, John C. and Gentry, James W.,** "Investigation of the Preference-Reversal Phenomenon in a New Product Introduction Task," *Journal of Applied Psychology*, No. 6, 1980, *65*, 715–22.

**Munera, Hector A. and de Neufville, Richard,** "A Decision Analysis Model when the Substitution Principle is not Acceptable," presented to the First International Conference on Foundations of Utility and Risk Theory, Oslo, July 1982.

**Pommerehne, Werner W., Schneider, Friedrich and Zweifel, Peter,** "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, *72*, 569–74.

**Ranyard, Rob H.,** "Elimination by Aspects as a Decision Rule for Risky Choice," *Acta Psychologica*, August 1976, *40*, 299–310.

**Reilly, Robert J.,** "Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design," *American Economic Review*, June 1982, *72*, 576–84.

**Russell, Thomas and Thaler, Richard,** "The Relevance of Quasi-Rationality in Competitive Markets." Working Paper, Graduate School of Business and Public Administration, Cornell University, 1982.

**Schoemaker, Paul J. H. and Kunreuther, Howard C.,** "An Experimental Study of Insurance Decisions," *Journal of Risk and Insurance*, December 1979, *46*, 603–18.

**Simon, Herbert A.,** *Models of Man*, New York: Wiley, 1957.

**Slovic, Paul and Lichtenstein, Sarah,** "Relative Importance of Probabilities and Payoffs in Risk Taking," *Journal of Experimental Psychology Monograph*, November 1968, Pt. 2, *78*, 1–18.

______, **Fischhoff, Baruch and Lichtenstein, Sarah,** (1982a) "Response Mode, Framing, and Information-Processing Effects in Risk Assessment," in Robin Hogarth, ed., *New Directions for Methodology of Social and Behavioral Science: Question Framing and Response Consistency*, San Francisco: Jossey-Bass, 1982.

______, ______, **and** ______, (1982b) *Regulation of Risk: A Psychological Perspective*, Decision Research Report 82–11, Decision Research: Eugene, 1982.

——— **and MacPhillamy, Douglas,** "Dimensional Commensurability and Cue Utilization in Comparative Judgment," *Organizational Behavior and Human Performance*, April 1974, *11*, 172–94.

**Thaler, Richard,** "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, March 1980, *1*, 39–60.

**Tversky, Amos,** "Intransitivity of Preferences," *Psychological Review*, January 1969, *76*, 31–48.

——— **and Kahneman, Daniel,** "The Framing of Decisions and the Psychology of Choice," *Science*, January 1981, *211*, 453–58.

**Weber, Robert J.,** "The Allais Paradox, Dutch Auctions, and Alpha-Utility Theory," Working Paper, J.L. Kellogg Graduate School of Management, Northwestern University, September 1982.

**Wehrung, Donald, A., MacCrimmon, Kenneth, R. and Brothers, K. M.,** "Utility Measures: Comparisons of Domains, Stability, and Equivalence Procedures," Working Paper No. 603, University of British Columbia, February 1980.

# The Generalized Theory of Transfers and Welfare: Bilateral Transfers in a Multilateral World

*By* JAGDISH N. BHAGWATI, RICHARD A. BRECHER, AND TATSUO HATTA*

Paul Samuelson's (1952, 1954) classic papers on the transfer problem addressed two separate analytical issues: the "positive" effect of a transfer on the terms of trade; and the welfare effect of the transfer on the donor and the recipient.

Since then, a considerable body of literature has grown up on the positive analysis. While Samuelson (1954) himself had extended the $2\times2\times2$ free trade analysis to allow for tariffs and transport costs, subsequent writers have analyzed other extensions of the model: for example, to allow for nontraded goods as with leisure in Samuelson (1971); or general nontraded goods in John Chipman (1974) and Ronald Jones (1970, 1975).

Remarkably, however, the welfare analysis of transfers has not paralleled these developments. Since Wassily Leontief (1936) produced an example of immiserizing transfer from abroad and Samuelson (1947) argued that the example required market instability, the proposition that has monopolized attention has been that a transfer in the conventional $2\times2\times2$ model in its free trade version cannot immiserize the recipient or enrich the donor as long as world markets are stable (in the Walras sense). Interestingly, Samuelson (1954), who did extend the positive analysis to include tariffs, did not go on to ask whether immiserization of the transfer recipient (and hence symmetrically enrichment of the donor in a two-country model) could now arise consistent with market stability.

Recently, the welfare analysis of transfers has been extended in two different directions, both apparently unconnected, and both yielding the conclusion that transfers from abroad can be immiserizing (and that the donor may improve its welfare) despite market stability. One route to this conclusion has been the introduction of a third economic agent (or country) that is outside of the transfer process. In the Appendix of his 1960 paper analyzing the interaction between trade policy and income distribution, Harry Johnson discussed the possibility of welfare-paradoxical redistribution between two factor-income classes (capital and labor) in an open economy, thereby providing what can be interpreted as a treatment of the three-agent transfer problem for the case in which donor and recipient are both completely specialized in the ownership of a single different factor.[1] An independent analysis of the three-agent transfer problem, using a restrictive model with given endowments of goods and fixed coefficients in consumption, was also undertaken in an important paper by David Gale (1974).[2] Brecher and Bhagwati

---

*Bhagwati: Department of Economics, Columbia University, New York, NY 10027; Brecher: Department of Economics, Carleton University, Ottawa, ON K1S 5B6; Hatta: Department of Political Economy, The Johns Hopkins University, Baltimore, MD 21218. We thank the National Science Foundation, grant no. 5-24718, for partial financial support of the research underlying this paper. The paper was written when Brecher and Hatta were visiting Columbia University, 1981–82. Gratefully acknowledged are helpful comments and suggestions from John Chipman, Avinash Dixit, Jacques Drèze, Robert Feenstra, Jacob Frenkel, Ronald Jones, Murray Kemp, Andreu Mas-Colell, Michael Mussa, John Riley, Lars Svensson, and Robert Willig, from anonymous referees, and from seminar participants at Berkeley, Harvard, Minnesota, Rochester, Chicago and the University of California-Los Angeles.

[1] After the present paper was submitted for publication, and following its presentation at Rochester, our attention was drawn to this Appendix, which was noticed by a student of Ronald Jones. Subsequently, we learned from Makoto Yano that Motoshige Itoh had pointed out an important related paper by Ryuotaro Komiya and T. Shizuki (1967), whose condition (11) for the Johnson case anticipated our equation (12) below. We are grateful for having both of these references brought to our attention.

[2] Gale constructs an example in which the donor is enriched along with the recipient. Furthermore, this immediately implies that a reverse transfer will immis-

(1981) also independently pioneered this analysis in the context of a three-agent model where the recipient country is split into one subset of "national" factors and another of "foreign" factors, and the conditions for the immiserization of the national factors after receiving a transfer from abroad are analyzed and shown explicitly to be consistent with market stability.

Another route has been to consider transfers in the presence of exogenously specified domestic distortions. Thus, Brecher and Bhagwati (1982) have analyzed the case of a transfer in the presence of a production distortion in the recipient country and shown that the recipient can get immiserized despite market stability if the recipient's "overproduced" good is inferior in the donor's consumption. Hatta, in an early unpublished paper (1973a), has also demonstrated for a closed economy with constant-cost production that a transfer between two agents, when there is a distortionary wedge between producer and consumer prices, could immiserize the recipient consistent with market stability. Peter Diamond (1978) has also recently considered the welfare impact of transfers when a price distortion exists in an economy with convex technology, and he gives comparative-static results that are consistent with paradoxes.

This recent proliferation of paradoxical cases of immiserizing transfers (and enriching transfer payments) is reminiscent of the earlier multiplication of cases involving immiserizing growth, with Bhagwati's (1958) analysis of the case of a large country in free trade being followed by Harry Johnson's (1967) analysis of the case of a small country with a tariff.[3] The latter proliferation led to the generalized theory of immiserizing growth (Bhagwati, 1968b) whose major, influential proposition is that growth, in the presence of a distortion implying departure from full optimality, can be immiserizing since the primary gain from growth at optimal policies may be outweighed by an accentuation of the loss from the distortion vis-à-vis the optimal policies.

Can a similar, striking generalization be developed in regard to the transfer-induced paradoxes? It is the general conclusion of our analysis in this paper that, indeed, it can. We demonstrate that the phenomenon of immiserizing transfers from abroad (and the analytically symmetric phenomenon of enriching transfer payments) in the presence of market stability can arise only if there is a distortion characterizing the economy in question.

This general conclusion is critically dependent on our demonstration below that the three-agent case, which appears prima facie to involve no distortion while producing the noted paradoxes, is indeed characterized by what Bhagwati (1971) has called a *foreign* distortion, since the country is not using an optimal tariff. Moreover, the exercise of their joint monopoly power by the recipient and donor (viewed as members of a customs union) vis-à-vis the nonparticipant agent will be shown to eliminate the paradoxes in question.

Thus, in Section I, we develop the basic analysis of transfers when there are two economic agents (countries) engaged in the transfer process, but there is an added agent *outside* the transfer process so that we have a bilateral transfer in a multilateral context. Conditions are established for immiserization of the recipient, for enrichment of the donor and for the "double perversity" when these two paradoxical outcomes arise simultaneously. Economically intuitive explanations of these results are derived in a number of alternative ways.

---

erize the (new) recipient. A simple calculation, moreover, shows that the Gale example is Walras-stable. Gale's work has stimulated a number of papers, most of which assume fixed commodity endowments and/or fixed consumption coefficients. A notable exception is an analysis of the three-agent transfer problem by Makoto Yano (1981), who introduces substitutability in both production and consumption. (We are grateful to Ronald Jones and Peter Neary for drawing our attention to Yano's work, after the research for the present paper was virtually completed.) A fuller discussion of these and related papers, in an evidently growing and important literature, is provided by us (1982b) elsewhere.

[3] Robert Aumann and B. Peleg (1974) have rediscovered, in a restrictive model with no substitution in production, the immiserizing growth case of Bhagwati (1958). See also Bhagwati (1982).

In Section II, yet further intuition on these results, in consonance with the theory of immiserizing growth, is arrived at, and suitable geometry of the three-agent transfer problem is simultaneously developed. Importantly, the role of inferiority in consumption or inelastic foreign demand is established in making feasible the perverse outcomes, which are shown to involve a foreign distortion (correctable by a uniform optimal tariff policy applied jointly by the donor and the recipient against the nonparticipant). In turn, this establishes an interesting parallel between the conditions for the immiserizing-transfer paradox in the three-agent, foreign-distortion case and the conditions established in Bhagwati's (1958) immiserizing-growth case which also involves a similar foreign distortion (i.e., growth for a large country that is failing to use an optimal tariff because of its free trade policy).

Section III then presents the implications of our results for some important theoretical and policy problems in both international and closed-economy contexts.

## I. Transfers with Three Agents: Model and Analysis

We begin with a formal analysis of the three-agent transfer problem, drawing on duality theory in terms of compensated demand functions, which have been introduced into the welfare-theoretic analysis of international trade by Hatta (1973b, 1977), Hatta and Takashi Fukushima (1979), and most notably and comprehensively by Avinash Dixit and V. Norman (1980), although earlier applications such as indirect utility functions are to be found also in the work of Chipman (1972).

### A. *The Model*

Consider a world economy consisting of three countries: $\alpha$, $\beta$, and $\gamma$. (While the analysis is couched in terms of three countries, it is applicable immediately to a closed-economy context with three agents within the economy, or to a two-country international economy where one country is disaggregated into two groups as in Brecher-Bhagwati, 1981.) Each country produces and consumes two goods, $X$ and $Y$. Free trade and perfect competition prevail.

Now, suppose that country $\alpha$ makes a transfer to country $\gamma$. Country $\beta$ does not participate in the transfer process. We will call $\alpha$ the *donor*, $\gamma$ the *recipient*, and $\beta$ the *nonparticipant* "outside" country. The objective of the analysis will be to determine the effect of the transfer on the welfare levels of the three countries.

The following notation will be used in presenting our model:

$q$ = the relative price of good $X$,
$u^i$ = the welfare level of country $i$,
$T$ = the value of the transfer in terms of good $Y$,
$e^i(q,u^i)$ = the expenditure function of country $i$,
$r^i(q)$ = the revenue function of country $i$,
$x^i(q,u^i)$ = the compensated import-demand function for good $X$ by country $i$, for $i=\alpha,\beta,\gamma$.

We then define an *overspending function* $c^i$ as follows:

$$c^i(q,u^i)\equiv e^i(q,u^i)-r^i(q), \qquad i=\alpha,\beta,\gamma.$$

Evidently, the value of this overspending function represents the difference between the expenditure necessary to achieve the utility level $u^i$ when the goods-price ratio is $q$ *and* the revenue of the producers of country $i$ at the same price ratio. Thus, $c^i$ is the amount of added revenue (i.e., transfer income) that is necessary for this country to sustain $u^i$ when the price ratio is $q$.

Using this notation, we can write our model as follows:

$$(1)\qquad c^\alpha(q,u^\alpha)+T=0,$$

$$(2)\qquad c^\beta(q,u^\beta)=0,$$

$$(3)\qquad c^\gamma(q,u^\gamma)-T=0,$$

$$(4)\qquad x^\alpha(q,u^\alpha)+x^\beta(q,u^\beta)+x^\gamma(q,u^\gamma)=0.$$

This model of four equations contains four variables: $u^\alpha$, $u^\beta$, $u^\gamma$, and $q$. Equations (1)–(3) are the budget equations for the respective countries, while equation (4) is the market equilibrium condition for good $X$. (In view of Walras' Law, the market-clearing equation for good $Y$ has been omitted.)

### B. *Comparative Statics*

We now examine the impact of an exogenous increase in $T$ upon the variables of the model above. Throughout the paper, subscripts always indicate partial differentiation with respect to a particular variable; for example, $c_u^\alpha \equiv \partial c^\alpha/\partial u^\alpha$ and $x_q^\gamma \equiv \partial x^\gamma/\partial q$. The following theorem can now be derived.

THEOREM 1: *Assume (without loss of generality) that* $e_u^\alpha = e_u^\beta = e_u^\gamma = 1$ *initially; and let* $\Delta \equiv x^\alpha x_u^\alpha + x^\beta x_u^\beta + x^\gamma x_u^\gamma - x_q$, *where* $x_q \equiv x_q^\alpha + x_q^\beta + x_q^\gamma$. *Then*

$$(5)\quad dq/dT = (x_u^\gamma - x_u^\alpha)/\Delta,$$

$$(6)\quad du^\alpha/dT = \left[x_q - x^\beta(x_u^\beta - x_u^\gamma)\right]/\Delta,$$

$$(7)\quad du^\beta/dT = -\left[x^\beta(x_u^\gamma - x_u^\alpha)\right]/\Delta,$$

$$(8)\quad du^\gamma/dT = -\left[x_q - x^\beta(x_u^\beta - x_u^\alpha)\right]/\Delta.$$

PROOF:

Taking the total differential of (1) through (4), applying the assumptions of the theorem, and using the well-known property that $c_q^i = x^i$ (for $i = \alpha, \beta, \gamma$), we obtain

$$\begin{bmatrix} 1 & 0 & 0 & x^\alpha \\ 0 & 1 & 0 & x^\beta \\ 0 & 0 & 1 & x^\gamma \\ x_u^\alpha & x_u^\beta & x_u^\gamma & x_q \end{bmatrix} \begin{bmatrix} du^\alpha \\ du^\beta \\ du^\gamma \\ dq \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} dT.$$

Applying Cramer's Rule to this system and taking notice of (4), we immediately obtain the theorem.

It is readily shown that $\Delta$ equals (minus) the slope of the general equilibrium, excess-demand schedule of good $X$ for the world as a whole.[4] Thus, the Marshall-Lerner condition for (Walrasian) stability implies that $\Delta > 0$.

Note also that equations (5) and (7) yield $du^\beta/dT > 0$ if and only if $-x^\beta dq/dT > 0$. Thus, the welfare of the country not involved in the transfer improves if and only if the price of its export good goes up as a result of the transfer—as is indeed immediately evident.

### C. *Paradoxes: Enrichment of Donor and Immiserization of Recipient*

The welfare impacts of a transfer upon the donor and the recipient, however, are not as simple as this. In the remainder of this section, therefore, we will give various interpretations of Theorem 1, to shed more light on the conditions under which the paradoxes of immiserized recipient and enriched donor arise. Note immediately, however, that if either $x^\beta = 0$ or $(x_u^\beta - x_u^\gamma) = 0$, that is, if either $\beta$'s net trade is zero or $\beta$ and $\gamma$ share an identical marginal propensity to consume $X$,[5] the second term in the numerator of the right-hand side of (6) is zero. In this case, equation (6) reduces to $du^\alpha/dT = x_q/\Delta$, which is, of course, the familiar expression for the welfare effect on the donor in the two-country analysis. With $\Delta > 0$ and $x_q < 0$, $du^\alpha/dT$ must be negative, that is, the donor must be immiserized. When the only trade partner of the donor is the recipient, or when the recipient and the nonparticipant share an identical marginal propensity, therefore, the welfare impact on the donor is *as if* we were in a two-country world, and the donor paradox never arises. A symmetric conclusion can be derived for the welfare effect on the recipient from equation (8).

Generally, however, the second term in the numerator of the right-hand side of (6) or (8) can cause paradoxical welfare effects, and we have the following necessary conditions for the paradoxes:

---

[4] First use equations (1), (2), and (3) to write $u^\alpha = \nu^\alpha(q,T)$, $u^\beta = \nu^\beta(q)$ and $u^\gamma = \nu^\gamma(q,T)$, respectively. Then define $x^\alpha[q, \nu^\alpha(q,T)] + x^\beta[q, \nu^\beta(q)] + x^\gamma[q, \nu^\gamma(q,T)] \equiv \tilde{x}(q,T)$, which is the world's uncompensated excess-demand function for good $X$. Now, we have $\tilde{x}_q = x_q^\alpha + x_u^\alpha \nu_q^\alpha + x_q^\beta + x_u^\beta \nu_q^\beta + x_q^\gamma + x_u^\gamma \nu_q^\gamma = -x^\alpha x_u^\alpha - x^\beta x_u^\beta - x^\gamma x_u^\gamma + x_q \equiv -\Delta$.

[5] As may be readily verified, the marginal propensity to consume good $X$ in country $i$ equals $qx_u^i/e_u^i$ for $i = \alpha, \beta, \gamma$. Therefore, if $\beta$ and $\gamma$ share an identical marginal propensity to consume $X$, this implies that $x_u^\alpha = x_u^\gamma$ (recalling the normalization that $e_u^\alpha = e_u^\beta = e_u^\gamma = 1$).

$$(9) \quad \frac{du^{\alpha}}{dT} > 0 \text{ implies that } x^{\beta}\left(x_u^{\beta} - x_u^{\gamma}\right) < 0,$$

$$(10) \quad \frac{du^{\gamma}}{dT} < 0 \text{ implies that } x^{\beta}\left(x_u^{\beta} - x_u^{\alpha}\right) < 0.$$

In fact, (6) and (8) make it clear that when $x_q = 0$ (i.e., when substitution effects are assumed away), the second inequalities in (9) and (10) are not merely necessary but also sufficient conditions for the paradoxes. And it is equally clear that if $x_q$ is sufficiently negative (i.e., if $X$ and $Y$ are readily substitutable in production and consumption), the paradoxes are unlikely to occur.[6]

### D. *Decomposition of Welfare Changes*

In further understanding our results in Theorem 1, note first that the right-hand sides of (6)–(8) contain the $x_u^i$ terms for all countries $i$ $(=\alpha, \beta, \gamma)$, *except* for the one whose welfare is stated by the equation in question. Let us now examine why this curious fact holds; it leads us into an insightful way of looking at our results.[7]

For this purpose we may conceptually decompose the transfer from $\alpha$ to $\gamma$ into two stages. At the *first* stage, $\alpha$ gives transfers to both $\beta$ and $\gamma$ in proportion to their initial import demand for $X$.[8] At the *second* stage, $\beta$ gives $\gamma$ what it received from $\alpha$ in the first stage, with the final situation ending up therefore as equivalent to the actual transfer going exclusively from $\alpha$ to $\gamma$. The welfare effect on the donor $\alpha$ can then be decomposed into two effects corresponding to these two stages.

Rewriting (6), we have[9]

$$(11) \quad du/dT = x_q/\Delta - \frac{x^{\beta}}{\left(x^{\beta}+x^{\gamma}\right)} \frac{\left(x^{\beta}+x^{\gamma}\right)\left(x_u^{\beta}-x_u^{\gamma}\right)}{\Delta}.$$

Now it is possible to show that the first stage leads to the first term on the right-hand side of (11).[10] Making transfers to every other country in the world economy in proportion to its initial import demand for $X$ is therefore tantamount to making a transfer to the other country in a two-country context! This process therefore results in a negatively signed term; the paradox of donor enrichment cannot come from this stage.

On the other hand, the second stage leads to the second term on the right-hand side of (11).[11] The sign of this term depends exclu-

---

[6] This might explain why Gale (1974), who tried to construct an example of donor enrichment, wound up assuming fixed coefficients in consumption (with fixed coefficients in production also implied by his exchange model), and confessed his inability to admit "smooth preferences." Interestingly, the absence of smooth preferences also characterizes the examples that Gale attributes to other major mathematical economists such as Dreze and McFadden. Just recently, Daniel Leonard and Richard Manning (1982) provided a paradoxical example involving smooth preferences within an exchange model.

[7] In Section II and in fn. 15 below, we spell out an alternative way of seeing why the income terms of only the two "other" countries appear in equations (6) and (8).

[8] That is, when $\alpha$ gives out a transfer of one unit of $Y$, $\beta$ receives $x^{\beta}/(x^{\beta}+x^{\gamma})$ units of $Y$ and $\gamma$ receives $x^{\gamma}/(x^{\beta}+x^{\gamma})$ units of $Y$. If these ratios are positive, both countries receive positive amounts of transfer. If $x^{\beta}/(x^{\beta}+x^{\gamma})$ is negative, $\beta$ receives a negative transfer, viz, it gives a transfer of $-x^{\beta}/(x^{\beta}+x^{\gamma})$ to $\alpha$. In this situation, $\gamma$ receives one plus $-x^{\beta}/(x^{\beta}+x^{\gamma})$ units of $Y$, since $x^{\gamma}/(x^{\beta}+x^{\gamma}) = 1 - x^{\beta}/(x^{\beta}+x^{\gamma})$. Similarly, $\gamma$ gives a positive transfer to $\alpha$ when $x^{\gamma}/(x^{\beta}+x^{\gamma})$ is negative.

[9] The reader can similarly rewrite (8) for the welfare effect on the recipient, $\gamma$.

[10] To see this, appropriately exploit the essence of the result in (6). A unit transfer from $\alpha$ to $\beta$ implies

$$(a) \quad du^{\alpha}/dT^{\alpha\beta} = \left[x_q - x^{\gamma}\left(x_u^{\gamma} - x_u^{\beta}\right)\right]/\Delta,$$

and from $\alpha$ to $\gamma$ implies

$$(b) \quad du^{\alpha}/dT^{\alpha\gamma} = \left[x_q - x^{\beta}\left(x_u^{\beta} - x_u^{\gamma}\right)\right]/\Delta.$$

With actual transfers in the first stage divided according to the ratios $x^{\beta}/(x^{\beta}+x^{\gamma})$ and $x^{\gamma}/(x^{\beta}+x^{\gamma})$ between $\beta$ and $\gamma$, these ratios should be multiplied into (a) and (b), respectively. Adding the resulting equations yields $x_q/\Delta$.

[11] This readily follows from appropriately using the essence of the result in (7) and the fact that $x^{\beta}/(x^{\beta}+x^{\gamma})$ represents the share of the transfer $\beta$ received at the first stage and hands out to $\gamma$ in the second stage. Equation

sively on the direction of the price change caused by the second-stage transfer. We already know from (9) that $x^{\beta}(x_u^{\beta} - x_u^{\gamma}) < 0$ is a necessary condition for the paradox of donor enrichment, and why this occurs is readily seen from (11) and the second-stage argumentation.

### E. *Alternative Necessary Conditions for Paradoxes*

We now turn to an alternative, equally insightful way of looking at Theorem 1. We first establish a set of necessary conditions for the paradoxes of donor enrichment and recipient immiserization. Then, it will be shown how these conditions are also necessary for price amplification effects which further help to explain the paradoxical possibilities.

Take again the case of donor welfare, and apply the Slutsky equation to (6) to get[12]

$$(12)\quad \frac{du^{\alpha}}{dT} = \frac{\left(x_q^{\alpha} + x_q^{\gamma} + \tilde{x}_q^{\beta} + x^{\beta} x_u^{\gamma}\right)}{\Delta},$$

where $\tilde{x}^{\beta}(q)$ is the uncompensated import-demand function for country $\beta$. Now, given $\Delta > 0$, and assuming throughout the rest of this section without loss of generality that $x^{\beta} < 0$ (i.e., country $\beta$ exports good $X$), we see immediately that the donor can be enriched only if either $x_u^{\gamma} < 0$, or $\tilde{x}_q^{\beta} > 0$, or both. That is, if a transfer enriches donor $\alpha$, then *either* $X$ is an inferior good to the recipient $\gamma$ *or* the offer curve of the nonparticipant outside country $\beta$ is inelastic (such that the export supply of $X$ by $\beta$ falls as the relative price of $X$ rises).

Similarly, for the immiserization of the recipient, we must have $du^{\gamma}/dT < 0$, and this can be shown to imply that *either* $x_u^{\alpha} < 0$ *or* $\tilde{x}_q^{\beta} > 0$ or both.

To understand more fully why these conditions are necessary for the paradox of (say) donor enrichment, the Slutsky equation and (4) may be used straightforwardly to rewrite the stability condition as $\Delta = x^{\alpha}(x_u^{\alpha} - x_u^{\gamma}) - x_q^{\alpha} - x_q^{\gamma} - (\tilde{x}_q^{\beta} + x^{\beta} x_u^{\gamma}) > 0$. If $(\tilde{x}_q^{\beta} + x^{\beta} x_u^{\gamma}) > 0$—which can happen only if either $\tilde{x}_q^{\beta} > 0$ or $x_u^{\gamma} < 0$ (given still that $x^{\beta} < 0$)—$\Delta$ will be smaller than in the two-country case (in which $\tilde{x}_q^{\beta} = x^{\beta} = 0$), *ceteris paribus*. Therefore, the price change measured by equation (5) is amplified by the presence of the third (nonparticipant) country $\beta$. If this price-amplification effect applies to an improvement in the terms of trade for $\alpha$, the donor may be paradoxically enriched by the transfer, even though the (smaller) terms-of-trade improvement in the two-country case cannot be great enough for the paradox of donor enrichment. By similar reasoning,[13] if $\beta$'s offer curve is inelastic or good $X$ is inferior for $\alpha$, an amplified deterioration in $\gamma$'s terms of trade may be great enough for the paradox of recipient immiserization.

These necessary conditions for an international transfer paradox are, interestingly, analogous to those established by Bhagwati (1958) for immiserization due to domestic growth (in the form of factor-endowment expansion or technological improvement). As he showed, the paradoxical possibility of immiserizing growth requires that *either* growth be ultra-biased against production of the importable (i.e., the importable be an "inferior" good in production) *or* the foreign offer curve be inelastic. This analogy suggests immediately that, if immiserizing growth paradoxes are attributable to the presence of distortions, as shown in Bhagwati (1968b), it should be possible to interpret the present transfer analysis in the three-agent context also as one where the paradoxes of immiserized recipient and enriched donor arise only when a distortion is present. The distortion

---

(7) applies since, when $\beta$ makes a transfer to $\gamma$, the welfare effect on $\alpha$ is as if $\alpha$ is the nonparticipant, outside country; the resulting welfare impact on $\alpha$ per unit transfer from $\beta$ to $\gamma$ is

$$\frac{du^{\alpha}}{dT^{\beta\gamma}} = -\frac{x^{\alpha}\left(x_u^{\gamma} - x_u^{\beta}\right)}{\Delta} = -\frac{\left(x^{\beta} + x^{\gamma}\right)\left(x_u^{\beta} - x_u^{\gamma}\right)}{\Delta}.$$

[12] Substituting $\tilde{x}_q^{\beta} = x_q^{\beta} - x^{\beta} x_u^{\beta}$ into (6) and recalling $x_q = x_q^{\alpha} + x_q^{\beta} + x_q^{\gamma}$ yields (12).

[13] In this case, the Slutsky equation and (4) should be used to rewrite the stability condition as

$$\Delta = x^{\gamma}\left(x_u^{\gamma} - x_u^{\alpha}\right) - x_q^{\alpha} - x_q^{\gamma} - \tilde{x}_q^{\beta} - x_q^{\beta} x_u^{\alpha} > 0.$$

here, as in Bhagwati (1958), must again arise as a foreign distortion in the sense of Bhagwati (1971); that is, the failure to exploit monopoly power in trade. Indeed, this can be demonstrated, as in Section II below.

## II. Viewing Three-Agent Paradoxes as Resulting from Foreign Distortion

We now proceed to demonstrate that the perverse welfare responses to bilateral transfers in the multilateral framework of three agents are attributable to the presence of a foreign distortion, and that the introduction of a suitable optimal tariff that eliminates this distortion will rule out the paradoxes. We first demonstrate this geometrically, using a technique that is suitable for "large" (as well as "small") transfers.

### A. *The Geometry of the Free Trade Case*

We begin by illustrating in Figure 1 the possibility of a perverse welfare response to bilateral transfer in the three-agent case. For convenience of exposition without loss of generality, the diagram treats countries $\alpha$ and $\gamma$ as partners of a customs union engaged in (free) trade with country $\beta$. (This treatment takes on more than expositional importance in Part B below, when $\alpha$ and $\gamma$ uniformly impose an optimal tariff policy against $\beta$.) In the initial pretransfer equilibrium, the union produces on its production-possibility frontier $Q_yQ_x$ at point $Q$, consumes on its Scitovsky (1942) frontier $S_yS_x$ at point $S$, and trades with country $\beta$ from point $Q$ to point $S$ along the price line $QS$.[14] To avoid cluttering the diagram, we have not drawn country $\beta$'s offer curve, which starts at point $Q$ and passes through point $S$. For the sake of concreteness only, let country $\beta$ be again an exporter of commodity $X$ (i.e., $x^{\beta} < 0$) while country $\alpha$ imports this good (with $x^{\alpha} > 0$).

Now, with country $\alpha$ making a transfer to country $\gamma$, suppose that the former's terms of trade consequently improve because the marginal propensity to consume good $X$ is greater for country $\alpha$ than for its union partner. Figure 1 illustrates the borderline case in which the terms-of-trade improvement is exactly enough to leave country $\alpha$'s welfare unchanged, despite the transfer. The union shifts in production to point $Q'$ on curve $Q_yQ_x$, and moves in consumption to point $S'$ on curve $S_y'S_x'$, which is another Scitovsky frontier in the map corresponding to a constant level of country $\alpha$'s welfare. Country $\beta$'s offer curve (still not drawn) now starts at point $Q'$ and passes through point $S'$.

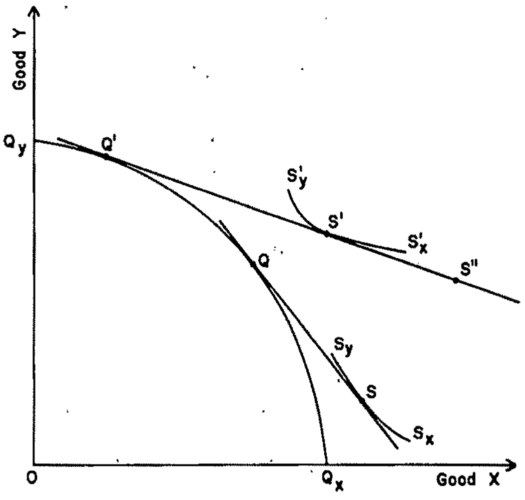


FIGURE 1

In Figure 1 as drawn, good $X$ is clearly inferior for the union as a whole. This inferiority, moreover, must characterize country $\gamma$ in particular, since country $\alpha$'s welfare is constant throughout the entire Scitovsky map. By contrast, no such inferiority would be implied if curve $S_y'S_x'$ were redrawn to touch line $Q'S'$ at point $S''$ (lying east of point $S$), while country $\beta$'s offer curve (not shown) were redrawn to pass through point $S''$ when starting at point $Q'$. In this alternative case, however, the offer curve of country $\beta$ must be inelastic, because a deterioration in this country's terms of trade is now associated with a rise in exports to the union. (These two alternative conditions are, of course, those already established in Section I,

---

[14] For further details on the use of production-possibility and Scitovsky frontiers corresponding to a pair of countries involved directly in a bilateral transfer, see Brecher and Bhagwati (1982).

Part E, above for the paradox of donor enrichment.)

By similar reasoning, an actual rise in country $\alpha$'s welfare, in response to a transfer to country $\gamma$, might occur provided that *either* good $X$ is inferior in the latter country *or* the offer curve of country $\beta$ is inelastic. This paradoxical possibility would occur in Figure 1 if country $\beta$'s offer curve (not drawn) were respecified to cross line $Q'S'$ southeast of the consumption point $S'$ (or alternatively $S''$) after starting at the production point $Q'$. In this case, after the transfer, there would remain a world excess supply of good $X$ if $q$ fell only enough to leave $u^{\alpha}$ constant at the initial (pretransfer) level. Thus, given stability, country $\alpha$'s terms of trade would ultimately have to improve still further, thereby leading to the paradox of donor enrichment.[15]

Essentially the same argument shows that country $\gamma$ might incur a welfare loss from receiving the transfer, provided that either good $X$ is inferior in country $\alpha$, or (as before) the offer curve of country $\beta$ is inelastic. If the required conditions for a perverse response in welfare hold simultaneously for both partners of the union, the transfer from country $\alpha$ to country $\gamma$ could raise the former's welfare and lower the latter's, implying a double perversity of outcomes.

### B. *Optimal Tariff Against a Nonparticipant Country*

Consider now the following alternative ways of demonstrating how the use of an optimal tariff by the union rules out the paradoxes at issue.

1. *A Geometric Analysis.* Consider the extension of the preceding analysis to the case where the union of $\alpha$ and $\gamma$ always maintains a uniform, optimal tariff vis-à-vis $\beta$, the nonparticipant country. Thus, for each value of the domestic goods-price ratio in Figure 2, the union adjusts the tariff to set the world-price ratio at the level consistent with the Robert Baldwin (1948) envelope $BE$, given the union's production-possibility frontier $Q_yQ_x$ and the offer curve $QG$ of country $\beta$. To avoid cluttering the diagram, we have drawn this offer curve with its origin in only one of the many possible positions. Alternatively, if this origin were placed at point $Q'$ (instead of $Q$) for example, the offer curve would touch curve $BE$ at point $S'$ (rather than $S$). Following a common convention, tariff revenues collected by each union member are returned to its consumers in lump sum fashion.

This optimal-tariff policy must result in the union consuming along its Baldwin envelope $BE$ of Figure 2 in equilibrium. Initially, the union produces on its production-possibility frontier $Q_yQ_x$ at point $Q$, consumes on its Scitovsky frontier $S_yS_x$ at point $S$, trades with country $\beta$ along the external-price line $QS$ from point $Q$ to point $S$, and imposes a tariff to create the proportional wedge between this price line and the (parallel) domestic-price lines (not drawn) tangent to curve $Q_yQ_x$ at point $Q$ and to curve $S_yS_x$ at point $S$.

---

[15] In determining whether there remains a world excess supply of good $X$ when $q$ is adjusted to keep $u^{\alpha}$ constant after the transfer, clearly the substitution effect but not the income effect plays a role in (the unchanged-welfare) country $\alpha$, whereas both of these effects are relevant in countries $\beta$ (as $u^{\beta}$ varies with $q$) and $\gamma$ (as $u^{\gamma}$ varies between Scitovsky frontiers). Thus, we have additional insight into why the income effects $x_u^{\beta}$ and $x_u^{\gamma}$ but not $x_u^{\alpha}$ enter the necessary condition (9). Similar reasoning sheds extra light on (10).

Two further remarks are in order. First, the use of the Scitovsky technique to analyze *welfare* changes here is only a natural counterpart to the two-stage derivation technique used by trade theorists in *positive* analyses. Thus, if the impact of a parametric change on the terms of trade is analyzed, one can hold the terms of trade constant, compute excess demand for one of the two goods, and then use the stability condition to determine the direction of the terms-of-trade change, a procedure introduced by Johnson, Mundell, and many other trade theorists in the 1950's. Identically, if one is interested instead in welfare change (of say the donor), one can equally hold welfare (of the donor) constant, compute excess demand, use the stability condition to determine the terms-of-trade change and then immediately the welfare change. This was, in fact, the procedure utilized in the analysis of immiserizing growth in Bhagwati (1958); and for the transfer problem, it leads naturally to the resurrection of the Scitovsky curves, as in the text above. Second, the use of Scitovsky curves in analyzing the transfer paradoxes was introduced earlier in Brecher and Bhagwati (1982). Whereas our Scitovsky technique easily handles large changes, the use of the two-stage holding-welfare-constant technique for small changes has also been introduced in Yano (1981) who credits Ronald Jones for the idea.

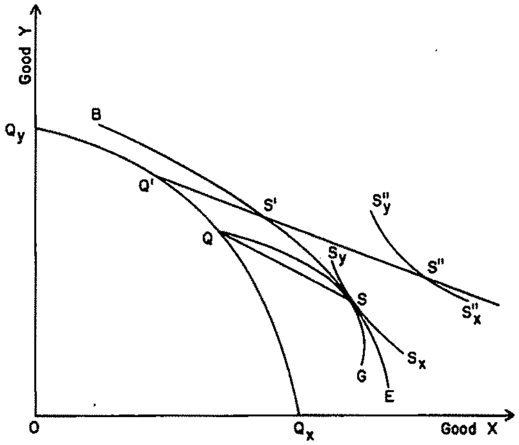


FIGURE 2

Now let the transfer take place from country $\alpha$ to country $\gamma$, and imagine what would hypothetically happen if the domestic relative price of good $X$ within the union fell exactly enough to leave the donor country $\alpha$'s welfare constant, despite the transfer. Under these circumstances, the union would move to point $Q'$ in production, trade along the external-price line $Q'S''$ with the rest of the world, and plan to consume at point $S''$ on the Scitovsky frontier $S_y''S_x''$ drawn for the initial level of country $\alpha$'s welfare. (For well-known reasons, the slope of curve $S_y''S_x''$ at point $S''$ equals the union's internal product-price ratio, given by the common slope of curves $Q_yQ_x$ and $BE$ at points $Q'$ and $S'$, respectively.) In this way, an excess demand (represented by the length $S'S''$) for good $X$ for the world as a whole would necessarily emerge,[16] and the relative price of this good would have to rise to clear world markets under stable conditions.

This implies that country $\alpha$, the donor, cannot enjoy enrichment since the initial fall in the price of $X$ which exactly offset the primary loss from the transfer would now be reduced, leaving $\alpha$ worse off. Similarly, we could establish that the recipient, $\gamma$, cannot be immiserized.

It follows, therefore, that the paradoxes of donor enrichment and recipient immiserization cannot arise if the union of the donor and the recipient follows the policy of adopting an optimal tariff that equates their domestic rates of (producer) transformation and (consumer) substitution to the foreign rate of transformation, such that $DRS^{\alpha} = DRS^{\gamma} = DRT^{\alpha} = DRT^{\gamma} = FRT$. Under free trade, however, the paradoxes become possible since the situation suffers from a foreign distortion such that $DRS^{\alpha} = DRS^{\gamma} = DRT^{\alpha} = DRT^{\gamma} \neq FRT$.

2. *Algebraic Analysis*: The preceding geometric analysis immediately suggests an approach to a formal proof of the proposition that the union of $\alpha$ and $\gamma$, utilizing an optimal tariff against $\beta$, would not admit of the paradoxes in question. Thus we should be able to show that, if such a tariff were in place, the paradoxes would be ruled out. This can indeed be done as follows.

Utilizing the model so far, we now distinguish between $q$ as the domestic relative price of $X$ and $p$ as the external price. Then, we write the foreign offer-curve function as $\bar{x}^{\beta}(p)$. Now, define the function $\rho$ by

$$(13)\qquad \rho(q,p) \equiv (q-p)\left[-\bar{x}^{\beta}(p)\right].$$

If $q$ and $p$ take on their equilibrium values, then $\rho(q,p)$ gives the tariff revenue of the union of $\alpha$ and $\gamma$. Also, let $p^*(q)$ represent the value of $p$ that maximizes $\rho(q,p)$, given $q$, and define the function $\rho^*$ by

$$(14)\qquad \rho^*(q) \equiv \rho\left[q, p^*(q)\right].$$

Now, if the international market is in equilibrium with $p = p^*(q)$, so that $\rho(q,p)$ takes on its maximized value $\rho^*(q)$ for the prevailing $q$, then it can be readily observed that the union must be operating on its Baldwin envelope at the point where $FRT$ equals that $q$. In this case, therefore, $q - p^*(q)$ is equal to the union's optimal tariff rate. Of course, there is no a priori guarantee that any arbitrary $q$ and the corresponding $p^*(q)$ are equilibrium values; in general, they are not. But the observation above implies that if the union always sets the tariff rate equal to $q - p^*(q)$ for the prevailing $q$, then this rate coincides with the optimal tariff rate when equilibrium is reached. The following result can now be established.

---

[16] Note that point $S''$ in consumption must lie outside the curve $BE$, assuming that good $Y$ is not sufficiently inferior to violate the Vanek (1965)-Bhagwati (1968a)-Kemp (1968) condition (discussed in more detail by us elsewhere, 1982a) for stability in the presence of tariffs. See also fn. 17 below.

LEMMA 1: *If* $p = p^*(q)$, *then* $(q-p)\tilde{x}^{\beta}_{p} = \tilde{x}^{\beta}$.

PROOF:

Recalling that the optimal tariff maximizes $\rho$ given $q$, we have $\partial \rho[q, p^*(q)]/\partial p = 0$, from which the lemma follows immediately.

Now, reformulate the model of this paper to allow for the union always imposing the tariff rate $q - p^*(q)$. Also assume that countries $\alpha$ and $\gamma$ collect tariff revenues equal to $[q - p^*(q)]x^{\alpha}(q, u^{\alpha})$ and $[q - p^*(q)]x^{\gamma}(q, u^{\gamma})$, respectively. (Thus, we implicitly assume that both union members import good $X$ from country $\beta$, although the results of the analysis would be essentially unaffected if one member received all of the tariff revenues because the partner imported nothing from $\beta$.) The overspending functions $c^{\alpha}$ and $c^{\gamma}$ are now given by

$$(15)\quad c^{\alpha}(q,u^{\alpha}) \equiv e^{\alpha}(q,u^{\alpha}) - r^{\alpha}(q) - [q-p^*(q)]x^{\alpha}(q,u^{\alpha}),$$

$$(16)\quad c^{\gamma}(q,u^{\gamma}) \equiv e^{\gamma}(q,u^{\gamma}) - r^{\gamma}(q) - [q-p^*(q)]x^{\gamma}(q,u^{\gamma}).$$

Then, our full revised model is given by

$$(17)\quad c^{\alpha}(q,u^{\alpha}) + T = 0,$$

$$(18)\quad c^{\gamma}(q,u^{\gamma}) - T = 0,$$

$$(19)\quad x^{\alpha}(q,u^{\alpha}) + \tilde{x}^{\beta}[p^*(q)] + x^{\gamma}(q,u^{\gamma}) = 0.$$

This three-equation model has three variables: $u^{\alpha}$, $u^{\gamma}$ and $q$.

The following theorem can be derived.

THEOREM 2: *If* (17)–(19) *hold and* (*by normalization*) *initially* $e^{\alpha}_{u} = e^{\gamma}_{u} = 1$, *then*

$$(20)\quad dq/dT = (x^{\gamma}_{u} - x^{\alpha}_{u})/\Delta^{t} c^{\alpha}_{u} c^{\gamma}_{u},$$

$$(21)\quad du^{\alpha}/dT = x^{t}_{q}/\Delta^{t} c^{\alpha}_{u} c^{\gamma}_{u},$$

$$(22)\quad du^{\gamma}/dT = -x^{t}_{q}/\Delta^{t} c^{\alpha}_{u} c^{\gamma}_{u},$$

*where* $\Delta^{t} \equiv -x^{t}_{q} + (c^{\alpha}_{q} x^{\alpha}_{u}/c^{\alpha}_{u}) + (c^{\gamma}_{q} x^{\gamma}_{u}/c^{\gamma}_{u})$ *and* $x^{t}_{q} \equiv x^{\alpha}_{q} + \tilde{x}^{\beta}_{p} p^*_{q} + x^{\gamma}_{q}$.

PROOF:

Taking the total differential of (17)–(19), we get

$$\begin{bmatrix} c^{\alpha}_{u} & 0 & c^{\alpha}_{q} \\ 0 & c^{\gamma}_{u} & c^{\gamma}_{q} \\ x^{\alpha}_{u} & x^{\gamma}_{u} & x^{t}_{q} \end{bmatrix} \begin{bmatrix} du^{\alpha} \\ du^{\gamma} \\ dq \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} dT.$$

We thus obtain

$$(23)\quad \frac{du^{\alpha}}{dT} = \frac{\left[c^{\gamma}_{u} x^{t}_{q} - (c^{\alpha}_{q} + c^{\gamma}_{q}) x^{\gamma}_{u}\right]}{\Delta^{t} c^{\alpha}_{u} c^{\gamma}_{u}}.$$

From (15) and (16), we get

$$c^{\alpha}_{q} + c^{\gamma}_{q} = p^*_{q}(x^{\alpha} + x^{\gamma}) - (q - p^*)(x^{\alpha}_{q} + x^{\gamma}_{q}),$$

and $$c^{\gamma}_{u} = 1 - (q - p^*)x^{\gamma}_{u}.$$

Thus, we have

$$\begin{aligned}(24)\quad & \left[c^{\gamma}_{u} x^{t}_{q} - (c^{\alpha}_{q} + c^{\gamma}_{q}) x^{\gamma}_{u}\right] \\ &= \left[x^{t}_{q} - (q - p^*)x^{\gamma}_{u}(x^{\alpha}_{q} + \tilde{x}^{\beta}_{p} p^*_{q} + x^{\gamma}_{q})\right] \\ &\quad - \left[p^*_{q}(x^{\alpha} + x^{\gamma}) - (q - p^*)(x^{\alpha}_{q} + x^{\gamma}_{q})\right] x^{\gamma}_{u} \\ &= x^{t}_{q} - x^{\gamma}_{u} p^*_{q}\left[(q - p^*)\tilde{x}^{\beta}_{p} - x^{\beta}\right] = x^{t}_{q},\end{aligned}$$

where the last expression follows from Lemma 1 and equation (19). Substituting (24) into (23) immediately yields (21). The other equations in Theorem 2 are derived similarly.

It is evident then from (21) and (22) that, with $\Delta^{\iota}c_u^{\alpha}c_u^{\gamma} > 0$ owing to market stability,[17] and with $x_q^{\iota} < 0$,[18] we necessarily get $du^{\alpha}/dT < 0$ and $du^{\gamma}/dT > 0$. That is, the donor must be immiserized and the recipient must be enriched. Paradoxes cannot arise.

3. *A General Proposition*. Now that we have demonstrated that the pursuit of an optimal tariff policy by the (union of the) donor and recipient jointly vis-à-vis the nonparticipant agent will rule out transfer paradoxes, in a Walras-stable context, we are able to see that the presence of a *suitably interpreted* (foreign) distortion is required in the three-agent case if the paradoxes are to arise.[19] At the same time, for the case of two agents, we know that *exogenously imposed* price distortions (for example, tax-cum-subsidies on production, consumption or trade) can also generate the transfer paradoxes (in the presence of inferior goods), as established by Brecher and Bhagwati (1982) and ourselves (1982a). We also know from the former paper, which analyzes transfer-induced distortions in the context of additionality requirements, and from the latter paper which analyses transfer-seeking *DUP* activities by domestic and foreign lobbyists, that *endogenous* (i.e., transfer-induced) distortions can also generate transfer paradoxes, consistent with Walrasian stability. We can therefore now state the following general proposition:

PROPOSITION: *The paradoxes of enriched donor and immiserized recipient cannot arise unless a distortion is present in the system.*

## III. Conclusion: Implications for Analytical and Policy Problems

The foregoing analysis has important implications in a number of areas of theoretical and policy concern.

*International*: (*i*) Our analysis of the three-agent problem does modify the earlier theoretical presumption against the possibility of stability-compatible paradoxes. (*ii*) Since, in the international context, reparations and aid are never given by one country to the "rest of the world," but are always bilateral transactions in a multilateral context, policymakers should be alert to the possibility that their intentions may be frustrated by paradoxical outcomes. (*iii*) As noted by Brecher and Bhagwati (1981), the three-agent transfer problem has an immediate counterpart in the analysis of customs unions with full mobility of factors within the union. Thus, for instance, it is possible for Italy to be immiserized within the EEC by receiving an aid inflow from the non-EEC world, under conditions established by us, consistent with market stability.

*Domestic*: (*i*) Internal redistribution from the rich to the poor may also be counterproductive under the conditions established here. Thus, if the poor receive the transfer from the rich while the not-so-poor outside group is a net exporter of food and the rich also have a lower marginal propensity to consume food than the not-so-poor, then we know that the conditions are satisfied to make it possible for the poor to be immiserized by receipt of the transfer. (*ii*) The three-agent analysis also brings into sharp focus problems raised by the "basic needs" prescription that the targeted poor be given purchasing power to buy their nourishment et al. If this purchasing power is taken from the rich, the nonparticipant not-so-poor may well find that their real income is diminished by a transfer-induced deterioration in their terms of trade (under an appropriate ranking of marginal propensities to consume), so that the poor become not-so-poor whereas the not-so-poor are reduced to the ranks of the poor! Indeed, our three-agent analysis similarly implies a certain caution in treating famine relief through transfers of purchasing power to the distressed income groups. Unless a similar security net is available elsewhere, you may then be pushing the malnourished not-so-obviously-starving poor (who are not receiving this purchasing power) below the line so that *they* are now ravaged by the famine.

---

[17] By reasoning similar to that of fn. 4 above, $\Delta^{\iota} = \tilde{x}_q^{\iota}$, where $\tilde{x}^{\iota}(q, T) \equiv x^{\alpha}[q, \nu^{\alpha}(q, T)] + \tilde{x}^{\beta}[p^*(q)] + x^{\gamma}[q, \nu^{\gamma}(q, T)]$, while the indirect utility functions $\nu^{\alpha}$ and $\nu^{\gamma}$ now come respectively from equations (17) and (18). Thus, Walrasian stability requires that $0 < \Delta^{\iota}$. Also, as explained by us elsewhere (1982a), $c_u^{\alpha}$ and $c_u^{\gamma}$ must both be positive to satisfy the Vanek-Bhagwati-Kemp condition assumed in fn. 16 above.

[18] As we move along the Baldwin envelope, $p_q^* > 0$ and $\tilde{x}_p^{\beta} < 0$ are well-known properties of economic efficiency. Thus, $x_q^{\alpha} + \tilde{x}_p^{\beta}p_q^* + x_q^{\gamma} < 0$.

[19] While our analysis can be viewed therefore as essentially providing a *conceptual* way of integrating the many-agent problem into the theory of distortions and welfare, our result on optimal tariffs also has direct *policy* relevance. For example, if there are two income classes, rich and poor, in an open economy with monopoly power in trade, we can definitely assert now that, if an optimal tariff is always in place, redistribution from the rich to the poor will not lead to paradoxical results.

*The Invisible Shakedown*: Our analysis also suggests a generalization of the idea underlying Gale's (1974) example where both the donor and the recipient are enriched by a transfer, at the expense of the nonparticipant outside agent. What is implied here is a seemingly innocuous process that involves enrichment at the expense of an unsuspecting agent. Through this process, the outside agent is hurt, for the benefit of the transfer-process agents, in a fashion that is by no means perceived as such, unlike in overt and visible instances such as where an optimal tariff may be levied against that agent.

Gale's example is, however, only one such instance: where the transfer is between the two agents ($\alpha$ and $\gamma$) with the third agent ($\beta$) remaining outside of the transfer process. But it is easy to see that one of the two agents (say $\alpha$) could equally exploit the third agent ($\beta$) by making a *direct* transfer to it, immiserizing it while enriching itself (and even, if need be, the other agent, $\gamma$): the conditions for this being readily established from equations (6)–(8) above. This is a clear case where a gift horse does need to be looked at in the mouth since, to mix metaphors ever so slightly, it turns out to be a Trojan horse.

The class of cases where (seemingly innocuous) transfers can improve the donor's welfare at the expense of either the direct recipient or an agent outside of the transfer process, or both, may then be christened generically as phenomena involving an *Invisible Shakedown*.

## REFERENCES

**Aumann, R. J. and Peleg, B.**, "A Note on Gale's Example," *Journal of Mathematical Economics*, August 1974, *2*, 209–11.

**Baldwin, Robert E.**, "Equilibrium in International Trade: A Diagrammatic Analysis," *Quarterly Journal of Economics*, November 1948, *62*, 748–62.

**Bhagwati, Jagdish N.**, "Immiserizing Growth: A Geometrical Note," *Review of Economic Studies*, June 1958, *25*, 201–05.

______, (1968a) "Gains from Trade Once Again," *Oxford Economic Papers*, July 1968, *20*, 137–48.

______, (1968b) "Distortions and Immiserizing Growth: A Generalization," *Review of Economic Studies*, October 1968, *35*, 481–85.

______, "The Generalized Theory of Distortions and Welfare," in his et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam: North-Holland, 1971.

______, "Immiserizing Growth and Negative Shadow Factor Prices: A Comment on Aumann and Peleg," mimeo., May 1982.

______, **Brecher, Richard A. and Hatta, Tatsuo** (1982a), "The Generalized Theory of Transfers and Welfare (II): Exogenous (Policy-imposed) and Endogenous (Transfer-induced) Distortions," mimeo., July 1982.

______, ______, **and** ______, (1982b), "The Paradoxes of Immiserizing Growth and Donor-Enriching (Recipient-Immiserizing) Transfers: A Tale of Two Literatures," Paper No. 15, International Economics Research Center, Columbia University, November 1982.

**Brecher, Richard A. and Bhagwati, Jagdish N.**, "Foreign Ownership and the Theory of Trade and Welfare," *Journal of Political Economy*, June 1981, *89*, 497–511.

______ **and** ______, "Immiserizing Transfers from Abroad," *Journal of International Economics*, November 1982, *13*, 353–64.

**Chipman, John S.**, "The Theory of Exploitative Trade and Investment Policies," in Louis E. DiMarco, ed., *International Economics and Development*, New York: Academic Press, 1972.

______, "The Transfer Problem Once Again," in G. Horwich and Paul Samuelson, eds., *Trade, Stability and Macroeconomics: Essays in Honor of Lloyd A. Metzler*, New York: Academic Press, 1974.

**Diamond, Peter A.,** "Tax Incidence in a Two-Good Model," *Journal of Public Economics*, June 1978, *9*, 283–99.

**Dixit, Avinash and Norman, V.,** *The Theory of International Trade*, Cambridge: Cambridge University Press, 1980.

**Gale, David,** "Exchange Equilibrium and Coalitions: An Example," *Journal of Mathematical Economics*, March 1974, *1*, 63–66.

**Hatta, Tatsuo,** (1973a) "Compensation Rules in Multiple-Consumer Economies," The Johns Hopkins University, mimeo., 1973.

______, (1973b) "A Theory of Piecemeal Policy Recommendations," unpublished doctoral dissertation, The Johns Hopkins University, 1973.

______, "A Recommendation for a Better Tariff Structure," *Econometrica*, November 1977, *45*, 1859–69.

______ **and Fukushima, Takashi,** "The Welfare Effect of Tariff Rate Reductions in a Many Country World," *Journal of International Economics*, November 1979, *9*, 503–11.

**Johnson, Harry G.,** "Income Distribution, the Offer Curve and the Effects of Tariffs," *Manchester School of Economics*, September 1960, *28*, 223–42.

______, "The Possibility of Income Losses from Increased Efficiency or Factor Accumulation in the Presence of Tariffs," *Economic Journal*, March 1967, *77*, 151–54.

**Jones, Ronald,** "The Transfer Problem Revisited," *Economica*, May 1970, *37*, 178–84.

______, "Presumption and the Transfer Problem," *Journal of International Economics*, 1975, *5*, 263–74; reprinted in his *International Trade: Essays in Theory*, Amsterdam: North-Holland, 1979.

**Kemp, Murray C.,** "Some Issues in the Analysis of Trade Gains," *Oxford Economic Papers*, July 1968, *20*, 149–61.

**Komiya, Ryuotaro and Shizuki, T.,** "Transfer Payments and Income Distribution, *Manchester School of Economics*, September 1967, *35*, 245–55.

**Leonard, Daniel and Manning, Richard,** "Advantageous Reallocations: A Constructive Example," mimeo., 1982; *Journal of International Economics*, forthcoming.

**Leontief, Wassily,** "Note on the Pure Theory of Capital Transfer," in *Explorations in Economics: Notes and Essays Contributed in Honor of F. W. Taussig*, New York: McGraw Hill, 1936.

**Samuelson, Paul A.,** *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

______, "The Transfer Problem and Transport Costs: The Terms of Trade When Impediments Are Absent," *Economic Journal*, June 1952, *62*, 278–304.

______, "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments," *Economic Journal*, June 1954, *64*, 264–89.

______, "On the Trail of Conventional Beliefs about the Transfer Problem," in J. Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam: North-Holland, 1971, 327–54.

**Scitovsky, Tibor,** "A Reconsideration of the Theory of Tariffs," *Review of Economic Studies*, No. 2, 1942, *9*, 89–110.

**Vanek, Jaroslav,** *General Equilibrium of International Discrimination: The Case of Customs Unions*, Cambridge: Harvard University Press, 1965.

**Yano, Makoto,** "Welfare Aspects in Transfer Problem: On the Validity of the 'Neo-Orthodox' Presumptions," mimeo., December 1981 (rev. 1982); *Journal of International Economics*, forthcoming.

# Fisher's Paradox and the Theory of Interest

*By* JEFFREY CARMICHAEL AND PETER W. STEBBING*

Irving Fisher's *Theory of Interest* has proved to be a most durable and influential contribution to economic theory. A central element of Fisher's contribution is the Fisher hypothesis that, over the longer term, the real rate of interest is approximately constant, being determined largely by time preference, with movements in the nominal interest rate reflecting movements in the rate of inflation one-for-one.

The paradox of the Fisher hypothesis is that, despite its wide acceptance as the cornerstone of interest rate theory, most empirical evidence appears to be inconsistent with the hypothesis, at least in its strictest form. This empirical anomaly was noted by Fisher himself, who attributed it to money illusion. In the fifty years following the publication of the *Theory of Interest*, a considerable literature has evolved around this paradox.[1] The unifying theme of this literature is a common acceptance of the essence of Fisher's hypothesis. Most of the energy in this literature has been devoted to explaining systematic movements in the real rate of interest, and to providing reasons why the relationship between the nominal interest rate and inflation may only be approximate in the real world. This paper attempts to shed a different light on the paradox.

In this paper we argue that while Fisher's hypothesis involves a proposition about real capital, empirical tests of the hypothesis have used data about financial assets. The use of these data is not surprising, given the complete lack of adequate data about returns to capital.[2] However, using data about financial assets to test a theory about real assets involves some implicit assumptions about the margins of substitution between money and financial assets on the one hand, and financial assets and capital on the other. We suggest that when the nature of these two margins is considered more closely, the reluctance of the data to support Fisher's hypothesis becomes less paradoxical. Indeed, in contrast with the widely expressed views of the literature, we argue that there is a simple explanation of the empirical regularities that is in no way incompatible with the principles of rational behavior.

We suggest an alternative hypothesis that has its roots in the works of both James Tobin and John Maynard Keynes. We argue that, provided there is some regulation of interest rates (a minimum degree of regulation being the nonpayment of interest on money balances) and provided there is a relatively high degree of substitutability between money and financial assets, the Fisher hypothesis, as applied to financial assets, may be completely inverted. That is, in after-tax terms, we suggest that the nominal rate of interest on these assets may be approximately constant over the longer term, with the real rate of interest moving inversely one-for-one with the rate of inflation.

Section I presents a brief review of the theoretical justifications of both the Fisher and the inverted Fisher hypotheses. Since the hypotheses are mutually exclusive, the resolution of the issue should be empirical. Sec-

---

*Reserve Bank of Australia, 65 Martin Place, Box 3947 GPO, Sydney, N.S.W. 2001. This work evolved from a paper prepared for a Conference on the Effects of Inflation in Australia, at the Australian National University, November/December 1981. We are grateful to Thomas Cargill, Pravin Trivedi, Mathew Butlin, Bill Norton, an anonymous referee, and colleagues at the Reserve Bank for helpful comments on earlier versions of the paper. The views expressed herein and any remaining errors are our own and should not be attributed to any of the above or to our employer.

[1] See Richard Roll for a review of the literature up to the early 1970's. More recent studies include Eugene Fama (1975, 1977), Patrick Hess and James Bicksler, Michael Darby, Thomas Cargill, Charles Nelson and G. William Schwert, John Carlson, Kenneth Garbade and Paul Wachtel, Alexander Holmes and Myron Kwast, Kajal Lahiri, Maurice Levi and John Makin, Vito Tanzi, and Frederic Mishkin.

[2] Probably the best source of data about returns to capital comes from earning yield information on equities. These data approximate *ex ante* real returns under very restrictive assumptions.

tion II discusses the statistical properties of some empirical tests of the inverted Fisher hypothesis, and we show that the alternative hypothesis is easier to test than the conventional Fisher hypothesis. Section III provides some empirical results for the United States and Australia.

## I. The Theory of Interest

Despite the dominant place of Fisher's hypothesis in the accepted theory of interest, it typically receives only sketchy theoretical justification. In Fisher's analysis, the hypothesis rested on the influence that individual time preference was assumed to exert on the marginal productivity of capital. A more formal derivation of the proposition in an optimizing model was later provided by Miguel Sidrauski.[3]

There are, however, other influences on real returns. One important influence, pointed out by Tobin in his seminal paper on monetary growth, is the nonpayment of interest on money balances. This theme has been picked up by a number of authors in explaining the apparent nonconstancy of real interest rates. The crucial characteristic of money is that, by convention, its nominal rate of return is regulated to be zero at all times (ignoring for the moment any implicit payments on money). By definition, therefore, its real rate of return fluctuates inversely one-for-one with the rate of inflation; that is, the return on money is forced by regulation to behave in the exact opposite manner to that predicted by Fisher's hypothesis.

Another important influence on interest rates is taxes. Despite the obvious relevance of taxes, they appear to have been overlooked prior to Michael Darby's contribution.[4] A fundamental aspect of taxes, that has not generally been observed in empirical work on the topic, is the need to work directly with after-tax rates of interest. In a correctly specified model, it is after-tax rates of interest that enter individual budget constraints and determine behavior. This given, it is not sufficient to formulate hypotheses in pre-tax terms and then look for the effects of taxes in the estimated coefficients.

Since empirical tests of Fisher's hypothesis utilize data about financial assets (i.e., nominally denominated, interest-bearing claims), it is necessary to examine the implications of the three basic influences (time preference, regulation, and taxes) for the behavior of financial interest rates. Under existing tax laws, holders of financial assets earn a real rate of return, $r$, but pay tax at some rate, $\theta$, on nominal interest income $i$ $(i=r+\pi$, where $\pi$ is the rate of inflation). That is, they receive an after-tax real return on financial assets $r_N$, of

$$r_N=(1-\theta)r-\theta\pi; \tag{1}$$

Since money balances pay no nominal interest, the observed before- and after-tax real returns on money are both minus the rate of inflation. It is generally accepted, however, that money yields an implicit return in terms of liquidity and transactions services, as well as free banking services, customer gifts, and so on. Optimizing individuals will tend to hold money and financial assets to the point at which their after-tax real yields (explicit plus implicit) are equal. Thus we can define the equilibrium relationship between the two returns as

$$r_N=z-\pi, \tag{2}$$

where $z$ (identically equal to the after-tax nominal interest rate $i_N$) is the implicit marginal return on money or, equivalently, the premium that must be paid to holders of financial assets to compensate them for the loss of monetary services.

A similar set of equilibrium relationships should hold between financial assets and capital (and therefore between money and capital). If capital offers an after-tax real return of $\delta_N$, we can define

$$\delta_N=x+r_N, \tag{3}$$

where $x$ is the premium (or discount) that must be paid to induce individuals to hold capital. In this case, the premium is likely to be dominated by the fundamentally different risk characteristics of capital and financial assets.

---

[3] The constancy of the real rate of interest in Sidrauski's model is usually linked to his assumption of a fixed rate of time preference but, as shown by Carmichael, a fixed rate of time preference is necessary but not sufficient to generate the Fisher relationship in Sidrauski's model; the additional necessary conditions are quite restrictive.

[4] In a recent paper, Joe Peek found strong evidence supporting the inclusion of taxes in analysis of Fisher's hypothesis.

Neither $x$ nor $z$ are necessarily constant. If all three assets were gross substitutes, a rise (fall) in the after-tax real yield on one, with the yield on the second held constant, would, in the absence of corner solutions, force the yield on the third to rise (fall), but by less than the rise (fall) in the first. In other words, $x$ and $z$ bear part of the burden of adjustment; the extent of the burden depends on the relevant elasticities of substitution.

Suppose for the moment that time preference does in fact determine $\delta_N$ as an approximate constant, as suggested by Fisher. In this case, financial assets are influenced by the Fisher effect coming from capital and the exact opposite effect coming from money. For Fisher's hypothesis to be supported by the data it is necessary that $x$ be approximately constant and that changes in $z$ be approximately equal to changes in the rate of inflation. In other words, financial assets and capital must be very close substitutes at the margin while financial assets and money must be very poor substitutes at the margin.

Given the nature of the three assets, this configuration seems somewhat implausible. More plausible is a high degree of substitutability at the margin between money and financial assets. Apart from the medium of exchange functions, money and financial assets are almost identical: they are both nominal stores of value, they have very similar risk characteristics and so on. A very plausible alternative hypothesis is that $z$, rather than $x$, is approximately constant, with the implication that the after-tax nominal rate of interest on financial assets will be approximately constant, while the after-tax real rate will tend to move inversely one-for-one with the rate of inflation. We will refer to this as the "inverted Fisher hypothesis."[5]

Note that the inverted Fisher hypothesis for financial assets leaves open the question of the behavior of the after-tax real rate of return on capital. Fisher's hypothesis could still hold for capital if financial assets and capital were very poor substitutes at the margin.

We should also point out that the idea that the yield on financial assets may not respond to inflation because of the constant nominal yield on money dates back at least to Keynes.[6] According to Keynes, the premium $z$ is determined by liquidity preference and, while generally sensitive to economic conditions, is invariant to changes in inflation. Keynes went further in asserting that the rate of return on capital would be forced to conform to the inflation-induced movements in the yield on financial assets; this implies a high degree of substitutability between financial assets and capital as well as between financial assets and money, thereby ruling out the Fisher effect for capital as well as for financial assets.

Before proceeding to the empirical work, there are two points that we wish to make about the inverted Fisher hypothesis. First, since the central relationship is heavily dependent on regulation, we would expect it to be less likely to hold in deregulated markets. Second, we would only expect the inverted Fisher hypothesis to hold in countries with moderate or low inflation rates; by accentuating the cost of regulation, hyperinflation would tend to reduce greatly the substitutability between money and other financial assets.

## II. Testing the Inverted Fisher Hypothesis

The conventional Fisher hypothesis has proved to be extremely difficult to test. Correctly stated, the Fisher hypothesis relates the nominal rate of interest to the expected, rather than the actual, rate of inflation, as indeed the inverted Fisher hypothesis relates the after-tax expected real rate of interest to the expected rate of inflation. The main difficulty in testing the Fisher hypothesis has been ambiguities involved in constructing data about expected inflation, and the errors-in-variables problem that arises when the actual rate of inflation is used in its place. In this section, we show that these problems are not present in testing the inverted Fisher hypothesis.

---

[5] A more formal derivation of the inverted Fisher hypothesis in an optimizing model characterized by regulation and substitution can be found in our 1981 paper.

[6] For a good discussion of Keynes' argument, see Stephen LeRoy. We are grateful to a referee for suggesting this reference and the similarity between our work and that of Keynes.

We begin with two identities. The first defines the after-tax nominal interest rate as the after-tax real rate plus the inflation rate. The second defines the same relationships as holding between the mathematical expectations of the three variables involved:

$$i_N = r_N + \pi, \tag{4}$$

$$Ei_N = Er_N + E\pi, \tag{5}$$

where $E$ is the expectations operator. To these definitions we add two assumptions needed to test the inverted Fisher hypothesis. The first is that expectations are unbiased:[7]

$$\pi = E\pi + \varepsilon, \tag{6}$$

where $\varepsilon$ is a random variable with mean zero and variance $\sigma_\varepsilon^2$. The second assumption is that the after-tax nominal interest rate is known at all points in time; that is,[8]

$$Ei_N = i_N. \tag{7}$$

Combining (4)–(7) yields

$$r_N - Er_N = E\pi - \pi = -\varepsilon. \tag{8}$$

In other words, if the current $i_N$ is known, the actual deviation of the after-tax real rate from its mean will be minus the deviation of the actual inflation rate from its mean.

We now add the inverted Fisher hypothesis in its correct expectations form:

$$Er_N = \alpha_0 - E\pi + \xi, \tag{9}$$

or, equivalently,

$$i_N = \alpha_0 + \xi, \tag{10}$$

where $\alpha_0$ is the (assumed) constant premium on the financial asset, labeled $z$ in Section I, and $\xi$ is a random error with mean zero and variance $\sigma_\xi^2$. We assume $\xi$ and $\varepsilon$ to be independent.

The inverted Fisher hypothesis can be tested in any of four ways. First, since the after-tax nominal rate of interest is hypothesized to be a constant, it should be independent of both $E\pi$ and $\pi$. Thus, equation (10) can be estimated in either of the following forms:

$$i_N = \alpha_0 + \alpha_1 E\pi + \xi, \tag{10$'$}$$

$$i_N = \alpha_0 + \alpha_1' \pi + \xi. \tag{10$''$}$$

The hypothesis tests are $\alpha_1 = 0$ and $\alpha_1' = 0$, respectively.

Next, using equation (8) to eliminate the unobservable variable $Er_N$, we can write (9) in the testable form

$$r_N = \alpha_0 + \alpha_2 E\pi + \xi - \varepsilon, \tag{9$'$}$$

where $\xi$ and $\varepsilon$ are independent random variables by assumption and, from equation (6), $E\pi$ and $\varepsilon$ are also independent. The hypothesis test in this case is $\alpha_2 = -1$.

Finally, we can substitute equation (6) as well as (8) into (9) to eliminate the unobservable variable $E\pi$, to obtain the estimating equation

$$r_N = \alpha_0 + \alpha_2' \pi + \xi - (1 + \alpha_2')\varepsilon, \tag{9$''$}$$

where again, the hypothesis test is $\alpha_2' = -1$. Under this null hypothesis, the error term $\varepsilon$ disappears from the right-hand side of (9″), with the result that the equation can be estimated (under the null hypothesis) without an errors-in-variables problem. This makes (9″) an attractive estimating form since it does not require the use of arbitrary data about inflation expectations. Indeed the only maintained hypothesis about expectations is unbiasedness, which is widely accepted nowadays as a minimal requirement for any sensible expectations-generating mechanism. For the test to be powerful, however, it should be capable of rejecting the alternative, conventional Fisher hypothesis that $\alpha_2' = 0$. The problem is that under the alternative hypothesis $\varepsilon$ does not disappear from (9″), with the result that (under the alternative hypothesis) the estimate of $\alpha_2'$ will contain an errors-in-variables bias. There are two important qualifications that should be made about this problem. First, while the bias in $\alpha_2'$ under the alternative hypothesis is negative it should be relatively minor in magnitude.[9] Second, since equation (9′) does not involve an errors-in-variables problem under either hypothesis, it provides another means of separating the two. Of course (9′) has the drawback that it does require data about expectations. In the empirical work below we work predominantly with equation (9″), although some estimates of (9′) using constructed data about expectations for the United States are included as supporting evidence.

---

[7] Fama was the first to use this assumption in testing the conventional Fisher hypothesis.

[8] When the individual purchases a financial asset, he knows the nominal interest rate that he will receive. The assumption in equation (7) implies that the individual also knows his marginal tax rate.

## III. Some Empirical Evidence

In this section, we report the results of estimating equation (9″) for the United States and Australia. Apart from the obvious benefits of examining data for more than one country, we chose Australia for convenience since we had ready access to data.[10]

One qualification should be expressed about any test of either the Fisher hypothesis or the inverted Fisher hypothesis and this is that both hypotheses concern longer-run behavior. Any number of factors may cause short-run behavior to deviate from these longer-run forces. A complete model of interest rate behavior would need to incorporate these factors. At the same time, if the hypotheses have any relevance at all, it is not unreasonable to expect such fundamental bivariate relationships to be displayed by time-series data over a reasonable time horizon.

### A. *The United States*

In keeping with existing analyses of the U.S. data, we used quarterly data for the yield on three-month Treasury bills as a representative market interest rate series. The series was taken from the *Federal Reserve Bulletin* and represents the average yield on all accepted bids for the last issue of the quarter. The *ex post* after-tax real rate of interest was then calculated (according to the formula,[11] $r_N \equiv (1-\theta)i - \pi$) with the timing convention that $r_{Nt}$ is the actual after-tax real return on a three-month bill held from the beginning to the end of quarter $t$.[12] All rates of return were continuously compounded,[13] and, for convenience of interpretation, all variables were expressed as annual equivalents. The Consumer Price Index (*CPI*) was used to calculate the inflation rate, and the marginal tax series, $\theta$, was generously supplied by Vito Tanzi. This latter series, a marginal tax rate on interest income, was calculated as a weighted average of the marginal tax rates facing investors with different income levels; the series was calculated from actual tax return data. The analysis focuses on the sample period 1953:I–1978:IV; prior to 1953 interest rates were essentially pegged, further as pointed out by Eugene Fama, the *CPI* was substan-

---

[9] The bias is asymptotically equal to $-\sigma_\varepsilon^2/[\sigma_{E\pi}^2 + \sigma_\varepsilon^2]$. In other words, it will be small so long as the variance of expectational errors about inflation are small, relative to the variance of expected inflation itself. In the case of the U.S. data, for example, an optimal time-series predictor for expected inflation (described below) yields sample estimates of $\hat{\sigma}_\varepsilon^2 = .00096$ and $\hat{\sigma}_{E\pi}^2 = .00685$, and therefore of the bias $= -.123$.

[10] These data were constructed for a recent conference on the effects of inflation; see our 1981 paper.

[11] It is worth emphasizing again that the dependent variable suggested by theory is $r_N$ and not $r$. The few tests of the conventional Fisher hypothesis that have attempted to correct for taxes have done so by adjusting either the estimated coefficient on inflation or the inflation series itself, rather than the dependent variable. Since the theory relates $r_N$ to $\pi$, use of $r$ as the dependent variable would require adjusting both the constant and error terms for taxes, as well as adjusting the inflation series.

[12] Further details on the timing and sources of all data used are contained in a data appendix which is available from the authors on request.

[13] Continuous compounding avoids the need to allow for the cross-product term between $r_N$ and $\pi$ in the definition of the nominal interest rate.

TABLE 1—INVERTED FISHER HYPOTHESIS: U.S. THREE-MONTH TREASURY BILL RATE, 1953:I–1978:IV

| Equation | Constant $\alpha_0$ | Inflation $\alpha_2'$ | Time $\alpha_3$ | $R^2$ | $D$-$W$ | $\rho$ |
|---|---|---|---|---|---|---|
| 1.A | 1.190 | −1.022 | | .96 | 1.94 | .997 |
| | (2.49) | (47.06) | | | | |
| 1.B | | −1.024 | | .96 | 1.92 | |
| | | (47.56) | | | | |
| 1.C | | −1.021 | | .96 | 1.79 | |
| 1953:I–1971:II | | (39.62) | | | | |
| 1.D | | −1.012 | | .97 | 1.78 | |
| 1953:I–1965:IV | | (42.43) | | | | |
| 1.E | | −1.036 | | .94 | 1.95 | |
| 1966:I–1978:IV | | (28.53) | | | | |
| 1.F | .981 | −1.016 | .038 | .96 | 1.78 | .803 |
| | (2.90) | (44.65) | (6.11) | | | |
| 1.G | .976 | −1.010 | .037 | .97 | 1.80 | .799 |
| Expected Inflation | (2.93) | (49.19) | (6.07) | | | |

*Notes*: The dependent variable is $r_N$; $t$-statistics are in parentheses; $\rho$ is the value of the first-order autocorrelation coefficient; the value for $R^2$ is not strictly bounded by zero and unity in the equations without a constant term.

tially upgraded in 1953. The ending point of 1978:IV was dictated by the ending observation of the tax series.

The results of estimating equation (9″) by ordinary least squares are reported in Table 1. Equation 1.A gives the estimated coefficients of (9″) for the full data period. Analysis of the residuals of the *OLS* equation indicated the presence of first-order autocorrelation. The estimated coefficients in 1.A are corrected for first-order autocorrelation using the generalized Cochrane-Orcutt technique in the program AUTO developed by Adrian Pagan. The value of the estimated first-order autocorrelation coefficient, $\rho$, was very close to unity[14] which prompted us to estimate the equation in first-difference form. The results are reported in equation 1.B.

Equation 1.C reproduces equation 1.B over the shorter period 1953:I–1971:II used by Fama. Equations 1.D and 1.E reproduce the same equation over the two halves of the data period. Using 1.D and 1.E to test for structural change gives a test statistic of 0.29 compared with the critical value of $F(1,102) = 3.94$ at the 5 percent level of significance, thus supporting the null hypothesis of no structural change.

The inverted Fisher hypothesis, that $\alpha_2' = -1$, is strongly supported by the data according to equations 1.A and 1.B. The hypothesis $\alpha_2' = -1$ yields a $t$-statistic of 1.03 from equation 1.A and of 1.09 from equation 1.B. From the definition of real and nominal interest rates, $\alpha_2'$ insignificantly different from $-1$ implies $\alpha_1'$ in equation (9″) insignificantly different from zero. More importantly, the point estimates of $\alpha_2'$ are so close to minus unity that the deviations from minus unity are economically trivial. That is, according to these estimates, not only is inflation not reflected in the after-tax nominal interest rate one-for-one, it is not reflected at all. By extension, it would appear that movements in pre-tax interest rates have had more to do with movements in taxes than with movements in inflation. From Table 1 it is also clear that the inverted Fisher hypothesis is quite robust to the choice of time period, unlike the conventional Fisher hypothesis.[15]

---

[14] This result has been found elsewhere in the literature; see, for example, Holmes and Kwast.

[15] See for example, Cargill and Robert Meyer, and Cargill (1976); although for a counterview on the three-month Treasury bill rate, see Holmes and Kwast.

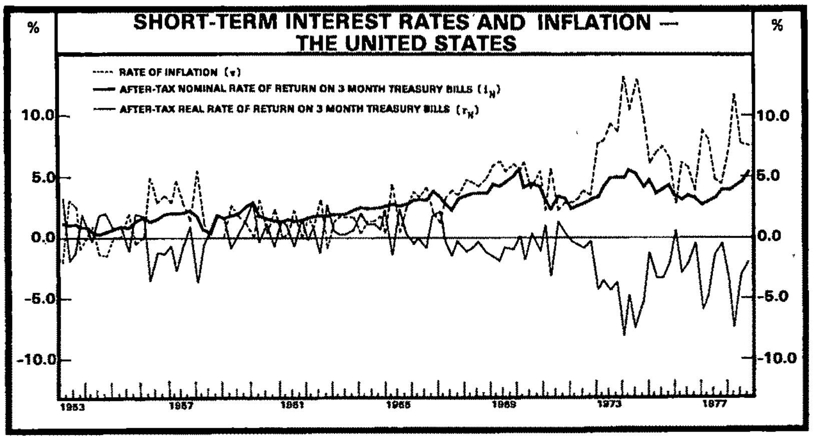


FIGURE 1

*Note*: Sample standard deviation of $i_N = 1.32$ and $\pi = 3.15$; sample correlation coefficient $i_N, \pi = .778$ and $r_N, \pi = -.931$.

The constancy of the after-tax nominal interest rate relative to inflation, and the inverse correlation between inflation and the after-tax real interest rate is shown in Figure 1 which graphs these three series. In Figure 1, the real rate $r_N$ is represented by the thin line, the nominal rate $i_N$ by the thick line, and the inflation rate by the broken line, $\pi$. Sample correlation coefficients are reported below the chart, along with the sample standard deviations of $i_N$ and $\pi$. The sample standard deviation of $i_N$ is 1.32, compared with 3.15 for $\pi$, a result that is consistent with the hypothesis of Section I.

Figure 1 does, however, show a discernible upward trend in $i_N$. Under the inverted Fisher hypothesis, $i_N$ is interpreted as the real risk premium on financial assets. Figure 1 suggests that this premium has drifted upwards over the last three decades by around 2 percentage points. The upward drift in $i_N$ is consistent with the widely held perception that implicit interest payments on money, particularly in the form of bank services and gifts, have risen over this period. There may, of course, be other possible explanations. In order to allow for possible bias in the coefficients as a result of the drift in $i_N$, we reestimated equation 1.A with a linear time trend; the results are reported in equation 1.F. While the coefficient on the time trend is significantly different from zero at all reasonable confidence levels, the properties of the equation are essentially unaltered.[16]

Equation 1.G reproduces 1.F, using an optimal time-series predictor to generate a series for expected inflation, in order to provide an additional test of the conventional Fisher hypothesis without an errors-in-variables bias. The model follows the work of Douglas Pearce who found evidence of a stationary first-order moving average process in the second difference of the *CPI* using monthly data from 1947–76. We estimated Pearce's model on monthly data from the beginning of 1947 to the end of 1978 and found the results to be compatible with Pearce's earlier results.[17] We used these estimates to generate nonoverlapping three-month forecasts of changes in the *CPI*. The estimates in equation 1.G are again completely consistent with the inverted Fisher hypothesis and provide no evidence of an errors-in-variables bias.

---

[16] The null hypothesis that $\alpha_2' = -1$ yields a $t$-statistic of .72 in equation 1.F.

[17] Pearce's model for the price level in period $t$, is $P_t = 2P_{t-1} - P_{t-2} + a_t + \gamma a_{t-1}$, where $P_t$ is the price level and $a_t$ is a random error. Using a rolling regression technique, Pearce found all estimates of $\gamma$ to lie between .71 and .76. Using the longer time period and assuming the process to be stationary (as suggested by Pearce), we estimated $\gamma = .75$ (with a standard error of .039).

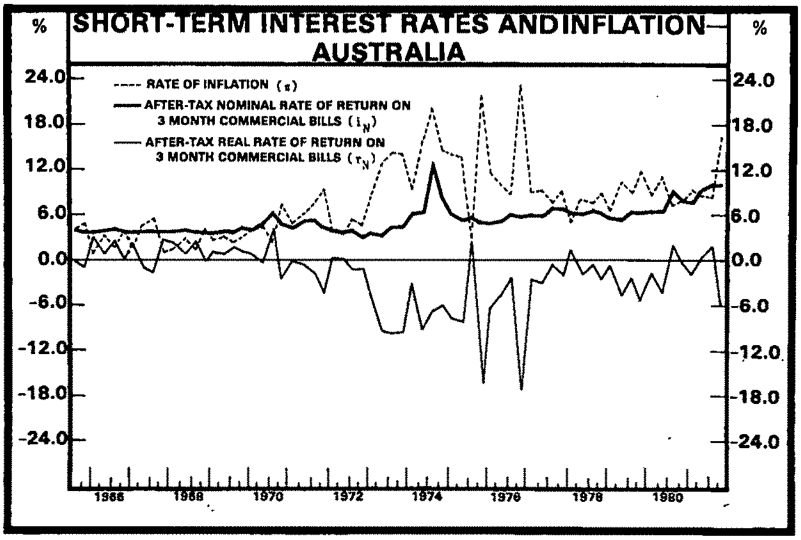


FIGURE 2

*Note*: Sample standard deviation of $i_N = 1.93$ and $\pi = 4.96$; sample correlation coefficient $i_N, \pi = .529$ and $r_N, \pi = -.923$.

## B. *Australia*

For Australia, we tested the inverted Fisher hypothesis on two interest rate series: a ninety-day bank accepted/endorsed commercial bill rate; and a five-year industrial debenture yield series. Both of these are market determined rates. The former is comparable with the data used for the United States, while the latter provides some evidence on a longer-maturity interest rate. We followed similar timing and construction conventions used for the U.S. data. The *CPI* was used for the rate of inflation and we used a tax rate constructed as the marginal tax rate facing a married person receiving average weekly earnings with three dependents.[18]

The starting points for the Australian study were dictated by the availability of the interest rate data. The earliest quarters for which data were available were 1965:II and 1962:I for the commercial bill and debenture rates, respectively. The end point for both series was 1981:IV.

Figures 2 and 3 show the after-tax nominal and real rates of interest with the inflation rate for the respective series. Once again, the sample correlations and sample standard deviations are reported below the charts. The results of estimating equation (9″) for the commercial bill rate are reported in Table 2, and for the debenture rate in Table 3.

As was the case for the U.S. Treasury bill rate, all equations showed evidence of first-order autocorrelation and were corrected using the generalized Cochrane-Orcutt procedure. Equation 2.E shows evidence of a secular upward drift in the premium on short-term interest rates in Australia. Equation 3.D suggests that this trend was not shared by longer rates, at least as measured by the industrial debenture yield.

In Table 2, equation 2.B reports estimates for (9″) over the shorter time period, 1968:II–1981:IV, which eliminates some less reliable observations prior to 1968.[19] Equa-

---

[18] Unfortunately, tax payments on actual interest income are not available for Australia. However, the entire tax schedule has drifted upwards over time so that tax payments at all levels of income have an essentially similar trend. Experiments with a range of tax rates yielded virtually identical results.

[19] Rates on the ninety-day commercial bill prior to January 1968 are based on a very small sample.

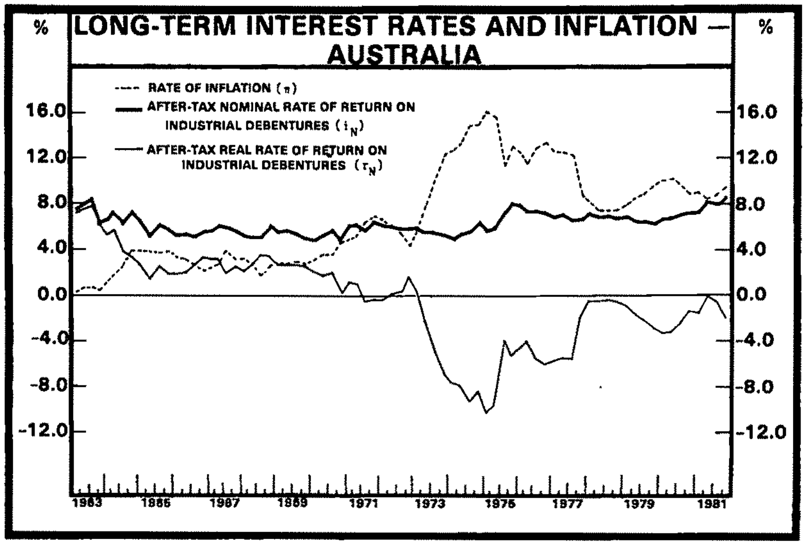


FIGURE 3

*Note*: Sample standard deviation of $i_N = 0.88$ and $\pi = 4.27$; sample correlation coefficient $i_N, \pi = .235$ and $r_N, \pi = -.978$.

TABLE 2—INVERTED FISHER HYPOTHESIS: AUSTRALIAN NINETY-DAY COMMERCIAL BILL RATE, 1965:III–1981:IV

| Equation | Constant $\alpha_0$ | Inflation $\alpha_2'$ | Time $\alpha_3$ | $R^2$ | D-W | $\rho$ |
|---|---|---|---|---|---|---|
| 2.A | 5.298 | −.979 | | .92 | 2.17 | .807 |
| | (7.42) | (27.56) | | | | |
| 2.B | 5.543 | −.973 | | .91 | 2.12 | .760 |
| 1968:II–1981:IV | (7.52) | (24.21) | | | | |
| 2.C | 4.258 | −1.015 | | .97 | 1.96 | .594 |
| 1965:III–1973:III | (14.21) | (24.97) | | | | |
| 2.D | 6.725 | −.983 | | .89 | 1.88 | .530 |
| 1973:IV–1981:IV | (7.88) | (17.40) | | | | |
| 2.E | 3.106 | −.982 | .072 | .93 | 1.98 | .552 |
| | (5.26) | (27.25) | (4.46) | | | |

*Notes*: See Table 1.

tions 2.C and 2.D and 3.B and 3.C report the results of dividing the relevant time periods in halves in order to test for parameter stability. The respective values for the $F$-statistic are 3.70 for the commercial bill rate and 1.58 for the debenture rate. A Chow test at the 5 percent level of significance rejects structural change for the latter, but not the former. However, a test for structural change in specific parameters indicates that $\alpha_2'$ is not significantly different in the two subperiods. The instability of the constant term is compatible with the significant time trend in equation 2.E.

The message from Tables 2 and 3 is a direct reinforcement of Table 1. The inverted Fisher hypothesis is strongly supported by evidence for both long and short rates for Australia; once again, this support appears to be robust across different time periods as

TABLE 3—INVERTED FISHER HYPOTHESIS: AUSTRALIAN FIVE-YEAR DEBENTURE YIELD, 1963:I–1981:IV

| Equation | Constant $\alpha_0$ | Inflation $\alpha_2'$ | Time $\alpha_3$ | $R^2$ | $D$-$W$ | $\rho$ |
|---|---|---|---|---|---|---|
| 3.A | 7.417 | −1.050 | | .99 | 2.29 | .909 |
| | (16.36) | (20.69) | | | | |
| 3.B | 6.746 | −1.136 | | .92 | 2.08 | .659 |
| 1963:I–1972:II | (15.37) | (9.75) | | | | |
| 3.C | 6.560 | −1.078 | | .98 | 1.97 | 1.000 |
| 1972:III–1981:IV | (13.16) | (21.43) | | | | |
| 3.D | 7.299 | −1.061 | .024 | .99 | 2.28 | .902 |
| | (14.51) | (17.97) | (.35) | | | |

*Notes*: See Table 1.

well as across different specifications.[20] The evidence for the longer rate is, if anything, more striking than that for the two short rates.

One of the most interesting features of the results in Tables 1, 2, and 3 is the high degree of explanation of movements in after-tax real rates of interest. The evidence of serially correlated residuals suggests immediately that variables have been omitted from the estimating equations. This is not unexpected, since the model is essentially one of longer-term behavior and has little to offer on short-term fluctuations in $i_N$. A more complete model would attempt to explain these fluctuations in terms of short-run supply-demand imbalances. The implication of the reported results, however, is that these short-run fluctuations are only a minor part of the story—movements in after-tax real interest rates are to a large extent explained simply by movements in the rate of inflation.[21]

If the strength of these results is surprising at first glance, it should not be. A negative relationship between the real rate of interest on financial assets and inflation has been noted elsewhere in the literature. For example, Frederic Mishkin found a strong negative correlation between the *ex post* real interest rate and inflation lagged one period. The relationship was stronger when it was adjusted for taxes. Others, such as Kajal Lahiri, John Carlson, and Pearce, have found a significant negative correlation between the real rate and expected inflation. Most have noted that nominal interest rates have risen by less than inflation. However, to the best of our knowledge, no one has considered an inverse one-for-one contemporaneous relationship between inflation and real rates as a logical theoretical possibility. It is only when this is taken as the starting point of the empirical analysis that the results stand out as sharply as they do here.

## IV. Conclusion

In this paper we have offered one explanation for what we have termed the Fisher paradox; namely the inconsistency between Fisher's hypothesis of a constant real interest rate and the empirical evidence. The fact that empirical tests of the hypothesis have been restricted to data about financial assets suggested the need to look more closely at the determinants of returns on these assets. In this respect, we argued that the margin of substitution between money and financial assets was likely to be the dominating influence. If correct, then not only would we not expect Fisher's hypothesis to be borne out by the available data, but there would be reasonable conditions under which we might expect the hypothesis to be completely inverted.

---

[20] To test for sensitivity of the results to particular data assumptions we also estimated equation (9″) using various alternatives including different tax variables, and seasonally adjusted data. The inverted Fisher hypothesis was robust across these alternative data definitions. A range of these equations is included in the data appendix available upon request.

[21] The remaining fluctuations are, of course, of considerable interest. We would expect them to be determined by a combination of real and financial shocks. On the role of aggregate supply shocks, see James Wilcox.

The inverted Fisher hypothesis, in its simplest form, is based on two assumptions: regulation in financial markets (in particular, regulation of nominal interest payments on money) and a relatively high degree of substitutability between regulated and nonregulated financial assets. The emphasis in the second assumption is on the term "relatively." In practical terms, a relatively high degree of substitutability translates into a low variance on after-tax nominal interest rates relative to the variance of inflation, a property which is supported by the data.

In testing this hypothesis, we considered data for the United States and Australia; for the latter, we tested both short-term and long-term interest rate data. The data were unable to reject the inverted Fisher hypothesis in any of the equations tested. The inverse one-for-one relationship between after-tax real interest rates and inflation received strong support from evidence in both countries. This support was robust across different time periods and maturities.

Contrary to some widely held beliefs, it would therefore seem that, as far as financial markets are concerned, the impact of inflation has fallen dominantly on real rates of return with little influence on nominal interest rates in either the short run or the long run. While this evidence provides one explanation of Fisher's paradox, it leaves open the possibility that Fisher's hypothesis may still hold for real assets such as capital. The large body of theoretical literature which seeks to explain movements in the real return to capital is still relevant to the margin of substitution between liquid assets as a group (money plus financial assets) and capital. Unfortunately, the lack of adequate data about *ex ante* returns to capital means that this issue must remain largely unresolved.

## REFERENCES

**Cargill, Thomas F.**, "Anticipated Price Changes and Nominal Interest Rates in the 1950's," *Review of Economics and Statistics*, August 1976, *58*, 364–67.

______, "Direct Evidence of the Darby Hypothesis for the United States," *Economic Inquiry*, January 1977, *15*, 132–34.

______ **and Meyer, Robert A.**, "Interest Rates and Prices Since 1950," *International Economic Review*, June 1974, *15*, 458–71.

**Carlson, John A.**, "Short-Term Interest Rates as Predictors of Inflation: Comment," *American Economic Review*, June 1977, *67*, 469–75.

**Carmichael, Jeffrey**, "Money and Growth: Some Old Theorems from a New Perspective," *Economic Record*, December 1982, *58*, 386–94.

______ **and Stebbing, Peter W.**, "Some Macroeconomic Effects of the Interaction Between Inflation and Taxation," paper presented at a Conference on the Effects of Inflation in Australia, Australian National University, Canberra, November/December 1981.

**Darby, Michael R.**, "The Financial and Tax Effects of Monetary Policy on Interest Rates," *Economic Inquiry*, June 1975, *13*, 266–76.

**Fama, Eugene, F.**, "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

______, "Short-Term Interest Rates as Predictors of Inflation: The Message in the Entrails," *American Economic Review*, June 1977, *67*, 487–96.

**Fisher, Irving**, *The Theory of Interest*, New York: MacMillan, 1930.

**Garbade, Kenneth and Wachtel, Paul**, "Time Variation in the Relationship Between Inflation and Interest Rates," *Journal of Monetary Economics*, November 1978, *4*, 755–65.

**Hess, Patrick J. and Bicksler, James L.**, "Capital Asset Prices Versus Time Series Models as Predictors of Inflation: The Expected Real Rate of Interest and Market Efficiency," *Journal of Financial Economics*, December 1975, *2*, 341–60.

**Holmes, Alexander B. and Kwast, Myron L.,** "Interest Rates and Inflationary Expectations: Tests for Structural Change 1952–1976," *Journal of Finance*, June 1979, *34*, 733–41.

**Lahiri, Kajal,** "On the Constancy of Real Interest Rates," *Economics Letters*, No. 1, 1979, *3*, 45–48.

**LeRoy, Stephen F.,** "Interest Rates and the Inflation Premium," *Federal Reserve Bank of Kansas Review*, May 1973, 11–18.

**Levi, Maurice D. and Makin, John H.,** "Anticipated Inflation and Interest Rates: Further Interpretation of Findings on the Fisher Equation," *American Economic Review*, December 1978, *68*, 801–12.

**Keynes, J. M.,** *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.

**Mishkin, Frederic S.,** "The Real Interest Rate: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy*: *The Cost and Consequences of Inflation*, Autumn 1981, *15*, 151–200.

**Nelson, Charles R. and Schwert, G. William,** "Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis that the Real Rate of Interest is Constant," *American Economic Review*, June 1977, *67*, 478–86.

**Pagan, Adrian N.,** "A Generalized Approach to the Treatment of Autocorrelation," *Australian Economic Papers*, December 1974, *13*, 267–80.

**Pearce, Douglas K.,** "Comparing Survey and Rational Measures of Expected Inflation: Forecast Performance and Interest Rate Effects," *Journal of Money, Credit and Banking*, November 1979, *11*, 447–56.

**Peek, Joe,** "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review*, December 1982, *72*, 980–91.

**Roll, Richard,** "Interest Rates on Monetary Assets and Commodity Prices Index Changes," *Journal of Finance*, May 1972, *27*, 251–77.

**Sidrauski, Miguel,** "Rational Choice and Patterns of Growth in a Monetary Economy," *American Economic Review*, May 1967, *57*, 534–44.

**Tanzi, Vito,** "Inflationary Expectations, Economic Activity, Taxes and Interest Rates," *American Economic Review*, March 1980, *70*, 12–21.

**Tobin, James,** "Money and Economic Growth," *Econometrica*, October 1965, *33*, 671–84.

**Wilcox, James A.,** "Why Real Interest Rates Were So Low in the 1970's," *American Economic Review*, March 1983, *73*, 44–53.

# A Competitive Theory of Monopoly Unionism

*By* EDWARD P. LAZEAR*

Most analyses of labor unions take the union itself to be the basic unit of analysis. The union is often discussed as if it were animate; possessing a utility function and displaying behavior consistent with the maximization of that function. Although models of this sort are not void of insight, they are less then completely satisfactory in providing a theory of labor unions that is general, that offers a large number of implications, and that is internally consistent.[1]

This paper attempts to fill some of the vacuum. The starting point is to take as the basic unit of analysis the profit-maximizing firms in which workers are employed, and the individual utility-maximizing workers themselves. The goal is to provide a model that predicts in which industries, occupations, and time periods unions are most likely to evolve. The analysis, albeit somewhat speculative, produces implications for the relationship between union wages, nonunion wages, and the proportion of workers in the union, and attempts to explain a number of institutions that seem to be quite specific to the union sector. The result is that there is a queue for union jobs but the nonunion sector always clears in the sense that supply equals demand. The union takes into account the effect of its wage choice on the resulting wage in the nonunion sector. Further, and equally important, is that firms are not passive in this model. Firms may resist the union at a cost and it is the distribution of those costs across firms that generates a well-defined equilibrium. Finally, the analysis attempts to come to grips with two often bothersome issues: first, because firms must bear a cost in order to prevent unionization from occurring, it is easy for union and nonunion firms to coexist in the same competitive industry. The marginal union firm and marginal nonunion firm have the same level of profits. Second, young workers, who seem to be at a disadvantage in the union hierarchy, have little incentive to offer their services to employers at a lower price. The seniority structure of the union is incentive compatible and guarantees maximum expected lifetime wealth to all workers.[2]

The major implications of the model are:

1) Consistent with Marshall, inelasticity of product demand increases the likelihood that an industry or occupation is unionized.

2) Contrary to Marshall, inelasticity of demand for labor does not imply an increase in union power as measured by either an increase in union membership or the union/nonunion wage differential.

3) The more elastic is the supply of labor to an industry or occupation, the smaller is the probability that that industry or occupation is unionized.

4) Not surprisingly, an increase in the cost of running the union and enforcing wage demands lowers the probability that an industry is unionized. Less obvious is that the union/nonunion wage differential increases as those costs rise. As a corollary, anti-union legislation decreases the proportion of unionized workers, but increases the union/nonunion wage differential. It is also true that occupations with less permanently attached workforces are less likely to be unionized, but also have higher union/nonunion wage differentials.

5) The observed wage differential may overstate rather than understate the true effect of unionism on wage rates. Nor is there

---

*University of Chicago, Graduate School of Business, 1101 East 58th Street, Chicago, IL 60637, and National Bureau of Economic Research. I gratefully acknowledge the helpful comments of K. Abraham, G. Becker, R. Hall, V. Lazear, J. Mincer, P. Mueser, G. Neumann, P. Pashigian, M. Reder, S. Rosen, and R. Topel. Financial support was provided by the National Science Foundation.

[1] See my more technical piece, 1983, that fills in some of the more sketchy parts of this paper.

[2] An early model, which provides for an active role by the firm, is Sherwin Rosen (1969).

any straightforward connection between wage differentials, proportion in the union, and union power.

6) Featherbedding is both rent maximizing and efficient. Further, a union which can select quantity as well as price, selects the pre-union employment level.

7) Union workers take a larger part of their compensation as fringes than do nonunion workers.

8) Union workers tend to be older than nonunion workers and receive more of the benefits of the union. Still, young workers have no incentive to break the coalition.

9) Age-earnings profiles are flatter in unionized firms.

10) Union/nonunion wage differentials move countercyclically.

The structure is to consider first a one-period setting since results are easier to obtain in that context. The second half of the paper examines a multiperiod world and shows that the results of the one-period model carry over. The strategy is to define the relevant labor supply and then to maximize the utility of the representative individuals subject to constraints imposed by the market and by the firm's active resistance.

## I. The Model

Consider a world in which a worker lives one period and where there are $S$ competitive firms, each of which has a demand for labor given by $L = d(W)$. This is simply the standard demand curve derived from the firm's production function where $W$ is the wage rate and $L$ is the number of workers in the firm. There are $R(W)$ workers in the occupation or industry. Initially (relaxed below), assume that $R$ is fixed and inelastic. Also assume for the moment that demand for output is perfectly elastic so that product market effects can be ignored.[3]

### A. *The Opportunity Locus*

A "union" is defined as a collection of workers who band together to call out a union wage, $W_U$. The firm, faced with the union demand, can pay that wage and hire $d(W_U)$ workers or at some fixed cost, $C_i$, which varies across the $i$ firms, can defeat the union and pay wage $W_N$, hiring $d(W_N)$ workers. The $C_i$ term is intentionally vague. It can be thought of in many ways: for example, it may be the cost of employing enough "union busters" to defeat the union or it might consist of contributions to an employee benefit fund to appease the current work force. It may be viewed as the cost of strikes or other morale problems associated with attempting to employ nonunion labor and so allows for the possibility that the firm beats the union by hiring scabs. Alternatively, if unions have beneficial effects on productivity as Richard Freeman (1976) and Charles Brown and James Medoff (1978) have argued, $C_i$ is the cost of foregone productivity effects.

Let $C_i \sim g(C_i)$ with a distribution function $G(C_i)$. The union knows the distribution of the $C_i$, but only the firm knows its particular $C_i$.[4]

The firm has a standard profit function of the form $\Pi = \Pi(W)$ so that the firm's profits, if it hires union labor, are $\Pi(W_U)$ and if it hires nonunion labor, $\Pi(W_N) - C_i$. The firm chooses to fight the union if

$$\Pi(W_U) < \Pi(W_N) - C_i$$

or if

$$\Pi^*(W_U, W_N) \equiv \Pi(W_N) - \Pi(W_U) > C_i.$$

For any given $(W_U, W_N)$ combination, then $G(\Pi^*(W_U, W_N))$ of the firms will find it more profitable to be nonunion firms.[5] Therefore the demand for labor by nonunion firms is $S[G(\Pi^*[W_U, W_N])]\, d(W_N)$ and by union firms is $S[1-G(\Pi^*[W_U,W_N])]d(W_U)$. The market equilibrium condition is that demand for labor equal the supply of labor:

---

[3] The supply of labor is that pool over which employers view workers as perfect substitutes in production.

[4] If the union actually knew the $C_i$, then the optimal strategy would be to price discriminate. Each firm would face a wage rate that made it just indifferent between hiring union workers and opposing the union. In fact, strikes are one manifestation of the firm's attempt to convince workers that its $C_i$ is low, while workers attempt to convince the firm that its $C_i$ is high.

[5] This subsumes all bargaining problems. Thus, Melvin Reder's (1952) notions of fairness and Carl Steven's (1958) early description of bargaining, as well as more modern bargaining models (Henry Farber, 1978; Vincent Crawford, 1979; W. Atherton, 1973) are implied.

$$S[1-G(\Pi^*[W_U,W_N])]\,d(W_U) + S[G(\Pi^*[W_U,W_N])]\,d(W_N)=R,$$

or in per firm notation,

$$(1)\quad [1-G(\Pi^*[W_U,W_N])]\,d(W_U) + G(\Pi^*[W_U,W_N])\,d(W_N)=R/S.$$

Equation (1) defines the union's "opportunity locus." It has the following interpretation: the union may select any $W_U$ that it chooses. However, its selection of $W_U$ affects the number of firms that opt to oppose the union. This, in turn, determines the number of union jobs available. Since the rest of the workers flow over into the nonunion sector, the wage rate in that sector adjusts to make supply equal to demand. Equation (1), therefore, incorporates the notion that firms can oppose unions and also the notion that the nonunion wage and, therefore, the strength of the firm's opposition, depends upon the union wage.[6]

The union wage affects the nonunion wage in two ways: first, a higher union wage increases the proportion of firms that fight the union. This implies a larger number of workers must be employed in the nonunion sector, but it also implies that there will be a larger number of nonunion firms to employ those workers. Second, a higher union wage decreases the number of workers hired by each of the union firms, sending the spillover into the nonunion sector.

It can be shown that the basic shape of the opportunity locus is always as illustrated in Figure 1. It starts at $(W_C, W_C)$ where $W_C$, the competitive wage, is derived from $Sd(W_C)=R$. It is negatively sloped initially, positively sloped as the union wage rises, always below $W_C$ so that the nonunion wage is inferior to the pre-existing competitive equilibrium, and it asymptotes to $W_C$.

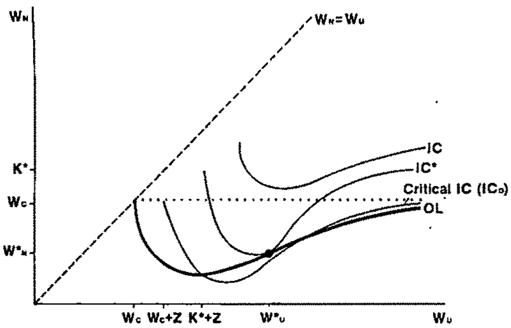


FIGURE 1

The intuition behind the asymptotic nature of the opportunity locus is clear. If the union chooses an extremely high wage, then all firms will opt to oppose the union. Under those circumstances, all firms and all workers are in the nonunion sector so the nonunion wage is simply the competitive wage. It is also clear that as the union wage approaches the competitive wage, the nonunion wage will also approach the competitive wage. The reason is that as the union wage shrinks to the competitive wage, union firms are demanding no less than their competitive share of labor and extracting workers from the nonunion sector in proportion to their numbers. This implies that workers remain in the nonunion sector in proportion to the number of firms there as well so the competitive wage must result. Similarly, whenever the union wage exceeds the competitive wage, union firms demand less than their competitive amount so the spillover of workers to the nonunion sector exceeds the number that those nonunion firms would have employed in competition. The nonunion wage therefore fails to bring about equilibrium.[7]

---

[6] One could also assume that the strength of the union's desire depends upon the difference between union and nonunion wages. This would make the distribution of $C_i$ a function of that wage differential. This possibility tremendously complicates the analysis and is ignored.

[7] The proof of these propositions, and of most others in this paper, is deleted. The complete analysis is contained in my 1983 article. Harry Johnson and Peter Miezkowski (1970) and W. Erwin Diewert (1974) offer alternative two-sector models of unionism. There, the union wage is exogenous and the purpose is to trace various wage choices through the rest of the economy in a general equilibrium framework. This model is interested in the way in which spillovers affect the *choice* of a union wage.

### B. *The Indifference Curves*

The $R$ workers who contemplate the union must take the opportunity locus as given. The group can select any union wage, but this implies a particular nonunion wage and a proportion of workers who will be able to locate union jobs. Following equation (1), it is clear that the probability that a given worker will be employed by a union firm and receive the union wage is

$$P \equiv S(1 - G[\Pi^*(W_U, W_N)])\, d(W_U)/R. \tag{2}$$

This is simply the number of union workers divided by the total labor supply.

In this one-period context, workers are identical and risk neutral *ex ante* so that each has the same objective function which is to maximize expected wealth. If it costs $Z$ dollars per union worker to administer the union, which may encompass hiring the union leader, striking, or other activities, then the worker's problem is

$$\underset{W_U}{\text{Max}}\, K = (W_U - Z) + (1 - P) W_N, \tag{3}$$

where $K$ is the level of expected utility, subject to the constraint implied by equation (1).[8]

The shape of the indifference curve is also illustrated in Figure 1. Those curves start at $W_N = K$ where $W_U = K + Z$ because workers only join the union if $W_U - Z > W_N$. They are U-shaped, but asymptote to $W_N = K$. The sense of the U-shaped indifference curve is this: workers would prefer to have both higher union wage and nonunion wage if nothing else were involved because they face some probability of finding themselves with a nonunion job as well as with a union job. But given that the union wage affects the probability of obtaining a union job, it is not always the case that workers view an increase in the union wage as a good. Initially, the value of an increase in the union wage outweighs the loss associated with the reduction in the probability of being unionized so indifference curves are negatively sloped. However, as the union wage rises and the probability of obtaining a union job get small, the benefits to increased union wages are swamped by the loss resulting from the decreased probability of obtaining a union job and workers view additional union wage increases as a bad, resulting in a U-shaped indifference curve.

It is necessary to define one more concept before proceeding to the comparative statics. Let us define the "critical indifference curve" as the one that yields the same utility as that derived in competitive equilibrium. The critical indifference curve is that curve that yields utility $K = W_C$, or from equation (3), it has the formula

$$W_N = [W_C - P(W_U - Z)]/(1 - P). \tag{4}$$

### C. *Equilibrium*

A union equilibrium exists if and only if the critical indifference curve crosses the opportunity locus. Since only points on the opportunity locus are feasible, a crossing implies there exists some feasible $(W_U, W_N, P)$ combination that makes workers better off than they would be in the competitive situation. The critical indifference curve often cuts the opportunity locus, but not always. For example, if $Z$, the cost of operating the union were very large, it is possible that no feasible combination exists which makes workers better off than they are in competition.

If a union equilibrium does exist, that is, if the critical indifference curve intersects the opportunity locus, the selection of the optimum union wage is given by the first-order conditions corresponding to equation (3) or by

---

[8] Note that this provides a well-defined notion of union membership. Before the election is held, all workers exert influence on the union. After the election, the union is observed to represent only those workers at union firms. Most of what has come before attempts to limit the relevant population to some subset of the labor supply. For example, Reder (1960), focuses on "present members." But this begs the question: present at what point in time? J. N. Dertouzos and John Pencavel (1980) implicitly confine their attention to *ex post* members as well and do not deal with spillover effects of the union wage on the nonunion wage.

$$(5a)\qquad \left.\frac{dW_N}{dW_U}\right|_{IC} = \left.\frac{dW_N}{dW_U}\right|_{0L};$$

$$(5b)\qquad [1-G(\Pi^*)]\,d(W_U)+G(\Pi^*)\times d(W_N)=R/S.$$

Equations (5) simply state that the optimum occurs at the tangency of the indifference curve with the opportunity locus.[9]

If a union equilibrium exists and if $Z$ is positive, then it is generally[10] the case that the tangency of $IC^*$ with the opportunity locus must be an interior one (see Figure 1). This implies that an occupation or industry is never entirely unionized. It is always better to allow those firms best able to resist the union to remain unorganized, rather than to choose such a low union wage as to make it unprofitable for any firm to oppose the union.

An important feature of this solution is that even if *ex post* losers would like to compete with *ex post* winners for union jobs, they are precluded from doing so. The firm can entertain these offers as a way to defeat the union, however, that carries with it the cost $C_i$ which is already accounted for in the analysis. Although the model allows for an active role by firms, it does not permit them to costlessly accept the labor offer of a nonunion worker. Even in the absence of NLRB rules, it seems reasonable to view the hiring of scabs as carrying some real costs, perhaps larger to some firms than others.

Note also that in equilibrium, the marginal union firm and marginal nonunion firm each earn the same level of profits. The high cost of labor in the union firm is counterbalanced by the cost of defeating the union in the nonunion firm. For the marginal nonunion firm, these costs are identical. Inframarginal rents are captured by scarce factors. Union and nonunion firms exist in the same product market, each earning the same level of profit, properly measured. There is no mechanism which results in one side driving the other side out of business.

Finally, although there is a queue for union jobs in the sense that all workers prefer union jobs to nonunion jobs, the nonunion sector is competitive and has the feature that wages are set in accordance with supply and demand. Workers may enter that sector freely and no artificial supply restrictions are required.

## II. Some Implications and Comparative Statics

### A. *The Supply of Labor*

In this section, the assumption that the supply of labor to the occupation is perfectly inelastic is relaxed. The conclusion is that as labor supply becomes more elastic, the probability that the industry is unionized falls. The logic is this: unionization must increase the expected wealth of individuals in that occupation. As such, more workers are induced to acquire the skills necessary to enter that occupation. In terms of Figure 1, this has the effect of shifting the opportunity locus downward and the critical indifference curve upward, reducing the probability of a union equilibrium. The intuition behind the downward shift in the opportunity locus is this: for a given union wage, a larger supply of labor results in a larger spillover of workers into the nonunion sector. In order to clear that market, the nonunion wage must fall by more than it would have had the supply of labor been fixed. Thus, $W_N$ is lower for any given $W_U$ and the opportunity locus shifts downward. Analogously, for the critical indifference curve, since the probability that a given worker will obtain a union job is lower the more workers there are in the industry, and, since an elastic labor supply implies more workers, at a given union wage the worker requires a higher nonunion wage to obtain the same level of expected utility. As a result, the critical indifference curve shifts upward.

At the extreme, the supply of labor to the occupation is perfectly elastic. Then, no union equilibrium could exist because raising the expected wage, even trivially, above the competitive wage brings about an infinitely sized labor force so that the probability of obtaining a union job falls to zero and all workers prefer the competitive situation. To the extent that the union can restrict entry into the occupation, this produces some inelasticity to the labor supply function and this case is then handled by the previous discussion.

---

[9] There may be more than on tangency; under these circumstances, it is necessary to search for a global maximum by comparing utility levels at all local maxima.

[10] This depends upon the distribution of $C_i$. The condition $G(0)=0$ is sufficient to guarantee the validity of the statements.

It might be reasonable to argue that the long-run supply of labor to an occupation is more inelastic than to an industry. Heterogeneity and talent across occupations make this a reasonable proposition. Historically, craft unions were organized before industrial unions and there is some evidence that the former have been more successful than the latter.[11]

### B. *The Demand for Labor*

Inelasticity of demand for labor does not necessarily increase the probability that a union will form. The reason is that although a smaller number of workers are displaced for a given increase in the union wage when the demand is inelastic, a given displacement reduces the nonunion wage by a greater amount when the demand is inelastic. Since a reduction in the nonunion wage implies a decrease in expected utility, the net effect of demand inelasticity is ambiguous. Effects on the nonunion sector were ignored by Marshall.

Surprisingly, what is important is the convexity of the demand curve. As the demand curve becomes more convex, a given increase in the union wage results in a smaller displacement of workers. At the same time, that given displacement into the nonunion sector reduces wages by a smaller amount when the demand curve is convex. As the result, the opportunity locus shifts upward and the critical indifference curve shifts downward.

This is no mere curiosity, for it carries with it the implication that as the demand for product becomes more inelastic, the probability that that industry is unionized rises. The reason is that the more inelastic is product demand, the more convex is the relevant labor demand curve.

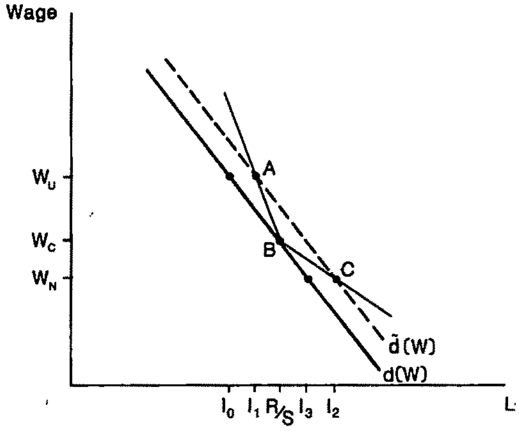


FIGURE 2

This is easily seen in Figure 2. Starting at the initial equilibrium, given by $(R/S, W_C)$, let the union raise the wage rate to $W_U$. The wage increase shifts average costs up for union firms by raising the price of labor and for nonunion firms, by raising the amount of resources expended by the marginal firm to defeat the union. If the demand for product is not perfectly elastic, then output price rises and the demand for labor shifts from $d(W)$ to $\tilde{d}(W)$ so that each union firm employs $l_1$ rather than $l_0$ of labor. Similarly, nonunion firms employ $l_2$ of labor and $W_N$ is the solution to equation (1). The relevant demand curve is then $AB$ and $BC$ for nonunion firms which adds "convexity" to the demand for labor. Further, the more inelastic is the demand for output, the larger is the increase in the price of the product associated with a given change in cost; this makes the demand curve even more convex. Since convexity increases the likelihood of a union equilibrium, a less elastic product demand results in a higher probability of a union.[12] Thus, Marshall's early statement is correct.

---

[11] See H. Gregg Lewis (1963) and Albert Rees (1962).

[12] There are additional second-order effects. There is no guarantee that the reduction in output by union firms plus the increase in output by nonunion firms will yield the net reduction in total output required as we move up the product demand curve. This will require a change in $S$, the number of firms in the industry.

### C. *The Cost of Operating the Union*

Above it was pointed out that as the cost of running the union, Z, rises, the likelihood of a union equilibrium diminishes. What is also true, however, is that, as that cost rises, the optimal union wage and resulting union/nonunion wage differential rises as well. Additionally, the proportion of workers with union jobs falls as operating costs rise.

In the context of Figure 1, the reason is as follows: an increase in $Z$ does not affect the opportunity locus since $Z$ does not enter equation (1). But, it can be shown that as $Z$ rises, the slope of the indifference curve through every point in the space falls. Flatter indifference curves imply that the tangency between an indifference curve and the opportunity locus lies farther to the right. This necessarily increases the optimal union wage. Further, since the union/nonunion wage differential can be measured by the vertical distance between the 45° line and the opportunity locus, and since the slope of the opportunity locus can be shown to be less than one, it is necessarily the case that the union/nonunion wage differential rises with $Z$ as well.

Consider, then, a law that makes it more costly to oppose the union. To the extent that this is interpretable as lowering Z, such a law increases the probability of unionization but decreases the optimal union wage and union/nonunion differential. Thus, right-to-work states should have a lower proportion of union workers, but the union/nonunion wage differential in those states should be larger than in non-right-to-work states. Incidentally, this yields the somewhat paradoxical result that pro-union legislation may actually increase efficiency. Although it raises the number of industries that are unionized, it reduces the amount of distortion in each so that the total effect is ambiguous.

### D. *Nationals and Locals*

The discussion so far has been in the context of a union selecting a wage for an entire industry or occupation, more like bargaining at the national than local level. Now suppose that each firm's $R/S$ workers act independently. The result is that the local union does not maximize the worker's expected wealth. The local union's behavior affects the market price and the aggregate probability that a worker obtains a union job, but the local union does not recognize that fact. Further, the local may establish a wage that is too low rather than too high. The analysis is lengthy and complicated and the reader is referred to my 1983 article for the complete discussion. The point, however, may be summarized in a few lines.

Locals and nationals differ primarily in two respects: first, a small local union might reasonably assume that its actions have no effect on the equilibrium nonunion wage. Second, an existing local ignores the affect that a change in that wage has on altering the probability that the firm will defeat the union. Because workers in each local do not take these effects into account, they may tend to favor unions to too great an extent. This yields the implication that the broader is the definition of the voting population, the more likely is the unit to recognize the market effects and, therefore, the less likely is the establishment of a union.

The obvious conclusion to be drawn here is that locals have an incentive to delegate wage-setting responsibilities to the nationals. Although any one local may be better off when locals adopt myopic behavior, the rest are sufficiently worse off to reduce the average wealth. To the extent that some lose and others gain, it may be necessary to couple the delegation of responsibility with transfers from one local to another. A centralized strike or pension fund, for example, which doles out benefits in accordance with some prearranged formula might be such a transfer mechanism.

### E. *Threat Effects and the Measurement of Union Power*

It is well known that the observed union differential may understate the true effect of the union on wage rates because nonunion firms, in attempting to discourage their workers from becoming unionized, pay more than the competitive wage.[13] The spillover of workers into the nonunion sector has the opposite effect. This is easily seen in the context of Figure 1. At any given union wage, the true amount by which unions raise wages is $W_U - W_C$ measured as the vertical distance between the 45° line and the horizontal line at $W_C$. The observed wage differential, however, is $W_U - W_N$, or the vertical distance between the 45° line and the opportunity locus. Since the opportunity locus is always below the competitive wage, the observed differential overstates rather than understates the true effect of unions on wages. Furthermore, since the difference between the competitive wage and the nonunion wage shrinks as the union wage approaches either the competitive wage or infinity, the overstatement of the true effect is smallest when the proportion of workers in the union is close to zero or one.

---

[13] See Rosen (1969) for a discussion.

It is important here to recognize that the assumption that $C_i$ does not take the form of an observed wage payment to nonunion workers is essential. In the standard threat-effect approach, the reason that the union/nonunion wage differential understates the true effect of unions is because the nonunion wage is brought up toward the union wage in an attempt to defeat the union. Lawrence Kahn and Freeman and Medoff provide evidence that threat effects are swamped by the wage depressant effect.

This raises a more fundamental issue. Given the logic of this paper, it is useful to inquire what we can learn from looking at wage differentials in relationship to other variables, especially the proportion unionized. The observed pair $(W_U, W_N)$ is the outcome of the union's optimization problem. Suppose that all unions faced the same opportunity locus and differed only in $Z$, the cost of running the union. Recall that as $Z$ rises, the optimal union wage rises and the proportion in the union falls, so occupations for which the costs of running a union are high have a high wage differential and few workers. From the low $P$, some might infer that the union is weak. From the high wage differential, others might infer that the union is strong. In some real sense, however, power is the same across occupations because the opportunity locus is unchanged. In another sense, the high $Z$ occupation is less powerful since its costs are higher and expected wealth lower. One might conclude that union power and proportion in the union are negatively related. Finally, a regression of the wage differential on the proportion in the union would yield a negative coefficient. If the opportunity locus shifts as well, the interpretation is even more confused. This suggests that inferences drawn with respect to union power on the basis of the classic study by Lewis or that by Rosen might be reexamined in this light.[14]

### F. *Quantity Restrictions and Price Discrimination*

It is well known that a monopolist with the power to do so should always set quantity equal to the efficient level.[15] Allowing the monopolistic union either to price discriminate or to set quantity as well as price through all or nothing contracts eliminates much of the inefficiency due to unions and results in a higher level of wealth for workers.

There are a number of ways by which a union can extract all the rent without affecting the quantity employed. One simple method is to use a two-part pricing scheme. The union wage is set equal to the competitive wage but the union requires a lump sum payment from the firm that is independent of the number of workers hired. The competitive wage guarantees that the firm hires the efficient number of workers and the lump sum extracts the rent in the most efficient manner. The finding by Freeman (1978) that union workers receive a higher proportion of compensation as fringes might be explained as part of the rent-extracting policy. An alternative strategy with equivalent results is to call out a higher union wage but require simultaneously that the firm employ the competitive number of workers. All rent can be extracted in that fashion as well without sacrificing efficiency.

---

[14] A similar point is made by Jacob Mincer (1981). His mechanism is somewhat different, however, since in his model workers queue as unemployed individuals.

[15] See Wassily Leontief (1946), William Fellner (1947), and Robert Hall and David Lilien (1979).

The firm is unlikely to sit by idly while all of this is occurring. The attempt by unions to charge a lump sum to firms provides incentives to merge. Similarly, at any wage rate higher than the competitive wage firms will resist the attempt by the union to require that the work force be held at its initial pre-union level. This has the appearance of "featherbedding," the requirement the firms hire more labor than they freely choose. This type of featherbedding is efficient, however. It provides that a Pareto optimum is reached since the competitive number of workers is employed. Furthermore, there is no offer that the firm can make to the union which will induce the union to eliminate this type of featherbedding, since it is an optimal rent extracting strategy as well. Inefficient featherbedding, where all could truly be made better off by its elimination, is discussed below.

### III. A Multiperiod Model

To this point the analysis has involved a one-period model where all workers are assumed to be identical. Yet, many of the interesting features of unions relate to worker differences and to how those differences are coordinated to produce a long-term relationship. Certainly of primary importance in this context is the issue of age and seniority. When workers vary by age and treatment of those workers is contingent upon age within the union, it seems natural that one group might compete with another. Structuring the union along seniority lines avoids this competition and as the result, the implications of the one-period model carry over to the multiperiod context.

An overlapping generations model captures the essence of life cycle differences within the union. Assume that workers live two periods and that there are $Y$ young persons born into the occupation each period so that $R=2Y$ is the labor supply. All workers are equally productive. How will union jobs be allocated by age? Suppose that all young workers had first claim to union jobs. If the occupation is less than completely unionized, then some older workers are left without union jobs. Those workers have an incentive to negotiate with the union employer and to offer services at a wage that exceeds the nonunion wage, but is less than the union wage. The union may or may not successfully defeat such attempts at "scabbing" by older workers. But the presence of such offers clearly raises the cost of operating the union and makes a union equilibrium less likely.

Consider the reverse situation. Let old workers have first claim to union jobs, but dole out the remaining union positions to the young workers. Young workers who care about expected lifetime wealth behave no differently. They remain loyal to the union even as nonunion workers because their entry into union jobs can be made contingent upon their cooperative behavior. Nor do old workers have any incentive to undermine the union since they are the individuals who reap the current benefits of such a scheme. As the result, the costs of operating the union are reduced which results in a higher expected lifetime wealth for all workers. Allocating the monopoly rents, first to the most senior workers and then to the junior workers, preserves incentive compatibility in a way that random assignments with respect to age does not.[16] The immediate implication is that union workers will tend to be older than nonunion workers.

A similar argument applies to retired workers as well. Since retired workers possess skills similar to those of the current work force, it may be important to discipline them. Some retired workers might desire to work occasionally and might offer their labor at a wage rate less than the union wage to union firms. If these workers can be punished for their anti-union behavior, then all workers can be made better off. Pensions may play an important role here. To the extent that old workers receive large pensions which are controlled at least in part by the union, scabbing may be discouraged. This suggests another reason why workers may receive a greater part of their lifetime compensation in the form of pensions than nonunion workers.

---

[16] A similar mechanism is discussed in my 1979 and 1981 studies.

"Inefficient featherbedding" can be rationalized along the same lines. The advantage to featherbedding is that the individual's payment is tied to the continued success of the union. A worker who is simply bought out of his union rights would otherwise have an incentive to undermine the union after his payoff was received.

Two additional points are worth mentioning: first, in a multiperiod context, $P$, the proportion of union workers should be reinterpreted as the proportion of each individual's lifetime spent in the union. Workers spend the first $(1-P)$ years of their lives as nonunion workers and the last $P$ years as union workers. Second, nonunion workers play a role even though they are not in the union. The situation where young workers queue up for union jobs and eventually become part of the union is probably best exemplified by the building trades. A worker may have to wait a number of years before he obtains a union card, and even then jobs may be assigned in an order related to seniority. If he does not obtain a union job on a given day, he may work in a nonunion job at a lower wage, but obviously cannot offer his services to the union employer. His willingness to comply rests on his knowledge that he will someday be the more-senior worker who reaps the returns of the union wage.

### A. *The Age Structure of the Union*

More formally, $P_0$ is the probability that an old worker obtains the union job and $P_Y$ is the probability that a young worker obtains a union job, then when jobs are allocated first to old workers and then to young workers the following expressions hold:

$$(6)\qquad P_0=\begin{cases}1 & \text{if } P\geqslant 1/2\\ 2P & \text{if } P<1/2,\end{cases}$$

where $\quad P\equiv S(1-G)\,d(W_U)/R.$

$$(7)\qquad P_Y=\begin{cases}2P-1 & \text{if } P\geqslant 1/2\\ 0 & \text{if } P<1/2.\end{cases}$$

If we define $\bar{A}_0$ as the average age of old workers and $\bar{A}_Y$ as the average age of young workers, then the average age of union and nonunion workers can be specified:

$$(8a)\qquad \bar{A}_U=\frac{P_0A_0}{2P}=\bar{A}_0 \quad \text{if} \quad P<1/2,$$

$$(8b)\qquad \bar{A}_U=\frac{\bar{A}_0}{2P}+\frac{(2P-1)}{2P}\bar{A}_Y \quad \text{if} \quad P\geqslant 1/2,$$

$$(8c)\qquad \bar{A}_N=\frac{(1-2P)\bar{A}_0}{2(1-P)}+\frac{\bar{A}_Y}{2(1-P)} \quad \text{if} \quad P<1/2,$$

$$(8d)\qquad \bar{A}_N=\frac{(2-2P)\bar{A}_Y}{2(1-P)}=\bar{A}_Y \quad \text{if} \quad P\geqslant 1/2.$$

Differentiation of equations (8) yields a number of interesting implications. First, as the proportion of the industry that is unionized rises, the average age of both union and nonunion workers falls. The intuition is that a larger proportion of union jobs induces the union to dip more deeply into the population of young workers. This lowers the average age among union workers. Additionally, since even among young workers, the oldest are taken first, increasing the size of the union sector leaves an even younger average group of nonunion workers. This is easily testable.

Additionally, one can derive a relationship between the difference between the average age of union and nonunion workers and the proportion of workers in the union. Differentiation of (8) with respect to $P$ yields the following result. The difference between the average age of union workers and that of nonunion workers within an occupation or industry first rises and then falls with increases in $P$. This too is easily tested.

### B. *Attrition*

Some occupations are characterized by higher turnover rates than others. There are a number of arguments that can be made that suggest that the costs of organization rise with the average rate of turnover within an occupation (obviously the turnover rate is not exogenous and independent of the resulting union and nonunion wages). To the extent that $Z$ rises, we have already derived a number of implications. First, industries where turnover rates are high are less likely to be unionized. Second, those high turnover rate industries that do become unionized should exhibit large union/nonunion wage differentials.

### C. *Multiperiod and Single Period Models*

The link between the overlapping-generation model in the simple one-period model is this: young workers in a multiperiod world behave exactly like single period wealth maximizers. Their critical indifference curves and their view of the opportunity locus are identical to those of individuals who live only one period. Old workers, however, may prefer a union equilibrium even when single period workers prefer the competitive equilibrium because the probability that an old worker obtains a union job exceeds $P$.

This poses no problem under a number of circumstances. The most straightforward resolution relies on the notion that the median voter is a young worker.[17] Although all young workers are not formal members of the union, their explicit or implicit acquiescence is necessary for the maintenance of a union equilibrium. Old workers would like to ignore the nonunion young workers in selecting $W_U$, but their persistence in following that strategy is not viable because young workers have an incentive to undermine such a union and replace it with a lifetime wealth-maximizing one. In fact, the young could pay old workers enough to induce the old workers to follow the broader lifetime wealth-maximizing policy. Entry fees and contributions by the current union workers to the retired workers' pension fund might be reinterpreted as a transfer of this sort.

Since young workers behave the same as workers who live for only a single period, and since a young worker is the marginal worker, the solution, which reflects the preference of the young worker, is identical to that in the single period world. Therefore, all results of that section hold.[18]

### D. *Age-Earnings Profiles*

Given that young workers are less likely to be union workers than are old workers within an occupation, there is a natural tendency for wages to grow over the life cycle as workers move from nonunion to union jobs. This is true even if productivity does not grow over the life cycle.

More important, within a union firm, age-earnings profiles will be flatter than they are in nonunion firms. The reason is that unions, in order to maintain stability, want the old rather than young workers to receive the benefits of the union. But in most industries and occupations, the firm, rather than the union, hires workers. A way by which the union can discourage the firm from hiring young workers is to overprice them relative to more experienced workers. This provides the union firm with incentives to hire the workers which the union would have selected.

The same argument applies to blacks and other groups that the union wants to keep out. The idea is to overprice black labor relative to white labor. Orley Ashenfelter (1972) finds support for this. Blacks are somewhat less likely to be in unions than whites, but those who are enjoy a smaller black-white wage differential.

### E. *Business Cycle*

Finally, Lewis reports that union workers do relatively better during cyclical downturns. The age-based incentive mechanism provides an explanation. Since incentive compatibility requires that old workers in unions receive benefits relative to young, seniority may be more important in determining layoff priority in union firms. As such, the average age of union workers relative to nonunion workers should rise during cyclical downturns and since age-earnings profiles are positively sloped, wage differentials will increase. This is easily tested by examining the way in which $\bar{A}_U - \bar{A}_N$ moves over the business cycle. Also, controlling for age should eliminate most of the countercyclical wage differential movement.

---

[17] See Farber for discussion of some basic aspects of union equilibrium in a median voter world.

[18] Again, see my 1983 study for the formal proofs.

## IV. Conclusion

This paper derives a large number of implications for the behavior and effects of unions. Although speculative, many are easily tested. The model is based on the standard assumptions in economic theory: firms maximize profit and workers maximize expected utility. The model is consistent in that it permits the coexistence of union and nonunion firms in the same product market and allows a queue for union jobs while nonunion wages are determined competitively. Incentive compatibility suggests that seniority plays an important role in the union and implications on the age structure and age-earnings profile in union and nonunion firms are derived.

## REFERENCES

**Ashenfelter, Orley,** "Racial Discrimination and Trade Unionism," *Journal of Political Economy*, May/June 1972, *80*, 435–64.

______, *Union Relative Wage Effects: New Evidence and a Survey of Their Implications for Wage Inflation*, Princeton: Princeton University, 1976.

**Atherton, W.,** *Theory of Union Bargaining Goals*, Princeton: Princeton University Press, 1973.

**Blau, Francine and Kahn, Lawrence,** "The Exit-Voice Model of Unionism: Some Further Evidence on Layoffs," unpublished paper, University of Illinois, 1980.

**Brown, Charles and Medoff, James,** "Trade Unions in the Production Process," *Journal of Political Economy*, June 1978, *86*, 355–78.

**Carlton, Dennis,** "Planning and Market Structure," in J. J. McCall, ed., *The Economics of Information and Uncertainty*, Chicago: University of Chicago Press, 1982.

**Carmichael, Lorne,** "Firm Specific Human Capital and Seniority Rules," Conference Paper No. 101, National Bureau of Economic Research, February 1981.

**Cohen, K. J. and Cyert, R. M.,** *Theory of the Firm: Resource Allocation in a Market Economy*, 2d edition, Englewood Cliffs: Prentice Hall, 1975.

**Crawford, Vincent P.,** "On Compulsory Arbitration Schemes," *Journal of Political Economy*, February 1979, *87*, 131–60.

**Dertouzos, J. N. and Pencavel, John H.,** "Wage Determinants Under Trade Unionism: The International Typographical Union," Stanford University, 1980.

**Diewert, W. E.,** "The Effects of Unionization on Wages and Employment: A General Equilibrium Analysis," *Economic Inquiry*, September 1974, *12*, 319–39.

**Dunlop, J. C.,** *Wage Determination Under Trade Unions*, New York: Macmillan, 1944.

**Epple, D., Hotz, V. J. and Zelenitz, A.,** "Employment Contracts, Risk Sharing and the Role of Unions," Working Paper No. 70-79-80, Carnegie-Mellon University, 1980.

**Farber, Henry,** (1978a) "Bargaining Theory, Wage Outcomes, and Occurrence of Strikes: An Econometric Analysis," *American Economic Review*, June 1978, *68*, 262–71.

______, (1978b) "Individual Preferences and Union Wage Determination: The Case of the United Mine Workers," *Journal of Political Economy*, October 1978, *86*, 923–42.

**Fellner, William,** "Prices and Wages Under Bilateral Monopoly," *Quarterly Journal of Economics*, August 1947, *61*, 503–32.

**Freeman, Richard B.,** "Individual Mobility and Union Voice in the Labor Market," *American Economic Review Proceedings*, May 1976, *66*, 361–68.

______, "The Effect of Trade Unionism on Fringe Benefits," Working Paper No. 292, National Bureau of Economic Research, October 1978.

______, "The Effect of Unionism on Worker Attachment to Firms," *Journal of Labor Research*, Spring 1980, *1*, 29–61.

**Hall, Robert E. and Lilien, David,** "Efficient Wage Bargains Under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

**Hicks, John R.,** *The Theory of Wages*, London: Macmillan, 1966.

**Johnson, Harry G. and Miezkowski, Peter,** "The Effects of Unionization on the Distribution of Income," *Quarterly Journal of Economics*, November 1970, *84*, 539–61.

**Kahn, Lawrence,** "Union Spillover Effects on Unorganized Labor Markets," *Journal of Human Resources*, Winter 1980, *15*, 87–98.

**Lazear, Edward,** "Why Is There Mandatory Retirement?," *Journal of Political Economy*, December 1979, *87*, 1261–84.

______, "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, *71*, 606–20.

______, "A Microeconomic Theory of Labor Unions," in Joseph Reid, ed., *New Approaches to Labor Unions, Research in Labor Economics*, Greenwich: JAI Press, 1983.

**Lewis, H. Gregg,** *Unionism and Relative Wages in the United States*, Chicago: University of Chicago Press, 1963.

**Leontief, Wassily,** "The Pure Theory of the Guaranteed Annual Wage Contract," *Journal of Political Economy*, February 1946, *54*, 76–79.

**Marshall, Alfred P.,** *Principles*, Book V, ch. VI, New York: Macmillan, 1961.

**Medoff, James,** "Layoffs and Alternatives under Trade Unions in United States Manufacturing," *American Economic Review*, June 1979, *69*, 380–95.

**Mincer, Jacob,** "The Economics of Wage Floors," Working Paper No. 804, National Bureau of Economic Research, November 1981.

**Pigou, A. C.,** *Economics of Welfare*, Book IV, ch. V, London: McNaught, 1920.

**Rees, Albert,** *The Economics of Trade Unions*, Chicago: University of Chicago Press, 1962.

**Reder, Melvin,** "A Theory of Union Wage Policy," *Review of Economics and Statistics*, February 1952, *34*, 34–45.

______, "Job Scarcity and the Nature of Union Power," *Industrial Labor Relations Review*, April 1960, *13*, 349–62.

**Rosen, Sherwin,** "Trade Union Power, Threat Effects and the Extent of Organization," *Review of Economics and Statistics*, April 1969, *36*, 185–96.

______, "Unionism and the Occupational Wage Structure in the United States," *International Economic Review*, June 1970, *11*, 269–86.

______, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, January/February 1974, *82*, 34–55.

**Ross, A. M.,** *Theory of Union Wage Policy*, Berkeley: University of California Press, 1948.

**Sayles, L. and Strauss, George,** *The Local Union*, New York: Harcourt, Brace and World, 1967.

**Stevens, Carl,** "On the Theory of Negotiation," *Quarterly Journal of Economics*, February 1958, *72*, 77–97.

**Stiglitz, Joseph,** "Utilitarianism and Horizontal Equity: The Case for Random Taxation," mimeo., 1980.

**Weinstein, Paul A.,** "The Featherbedding Problem," *American Economic Review Proceedings*, May 1964, *54*, 145–52.

# The Money Supply Announcements Puzzle: Review and Interpretation

*By* BRADFORD CORNELL*

The extent to which money affects the *ex ante* real interest rate is an issue debated for decades. The debate continues because empirical studies designed to measure the impact of monetary policy on the real rate are inconclusive. The problem is that money, prices, and interest rates, as well as Federal Reserve policy, are all endogenous. As a result, empirical tests of the relations among these variables are ambiguous without a structural model. Unfortunately, there is not yet a consensus on the appropriate structural model.

In this respect, studying the reaction of asset prices to money supply announcements presents a unique opportunity. Unlike the actual money supply, money supply announcements are exogenous. The money supply figure the Federal Reserve reports on Friday afternoon (previously Thursday afternoon) is the Fed's estimate of the money stock for the reserve week ending nine days earlier. At the time it is *announced*, the reported figure does not depend on the Fed policy, asset prices, or inflation. This means that if significant correlations are found between money supply announcements and changes in the prices of bonds and other assets, the direction of causation must run from the announcements to asset prices rather than vice versa. By examining these correlations for assets whose prices depend on the *ex ante* real interest rate, insight can be gained into the relation between the money stock and the real rate.

Another advantage in working with money supply announcements, rather than actual changes in the money supply, is that the observation interval can be greatly reduced. For example, studies by Frederic Mishkin (1981a,b, 1982) and Gerald Dwyer (1981) which examine the relation between interest rates and actual innovations in the money stock rely on monthly or quarterly data. The long observation interval is necessary to produce reliable estimates of the expected and unexpected change in money. On the other hand, studies of the relation between interest rates and *announced* innovations in the money supply use observation intervals of from two hours to one trading day. Unless the impact of monetary shocks on asset prices persists for a quarter or more, tests based on the shorter observation interval are more powerful.

Since the Federal Reserve's stated change in operating procedure on October 6, 1979, academic research on the impact of money supply announcements has increased. Between the beginning of 1981 and early 1983, more than a dozen papers on the subject appeared. In addition to presenting empirical results, these papers offer explanations for the positive correlation between announced innovations in money and changes in short-term interest rates observed by practitioners. Thomas Urich and Paul Wachtel (1981) extend the Keynesian model popularized in the financial press, my article (1983) stresses the impact of expected inflation, Donald Nichols, David Small, and Charles Webster (1983) point to the role of exogenous shocks to money demand, and Charles Engel and Jeffrey Frankel (1982) attempt to derive a general model incorporating all three effects. In this paper, a fourth link between money supply announcements and asset prices, based on variation in risk premiums, is proposed.

---

*Associate Professor of Finance, Graduate School of Management, University of California-Los Angeles, Los Angeles, CA 90024. I thank Jeff Frankel, Milton Friedman, Ron Masulis, Alan Shapiro, Lee Wakeman, and particularly Eugene Fama and Richard Roll for valuable comments on earlier versions of this paper. A preliminary version of this paper was presented at the Fall Academic Conference of the San Francisco Federal Reserve Bank.

Since each of the theories is designed to explain the positive correlation between announced money supply innovations and changes in short-term interest rates, tests using other assets are required to distinguish between the alternative hypotheses. These tests have a value beyond providing information on the market response to money supply announcements. The theories differ primarily in the way money is assumed to affect inflation, real activity, and the *ex ante* real rate. By comparing the theories, insight is gained into the relations among these variables. Thus studies of the reaction of asset prices to money supply announcements can provide partial answers to some of the most fundamental questions in monetary theory.

Unfortunately, a majority of the empirical papers do not concern themselves with this broader problem. Though there are exceptions, most of the papers focus on the relation between money supply announcements and the price movements of one specific asset such as Treasury bills, Treasury bonds, stock or foreign exchange. One purpose of this paper is to summarize and to extend the existing empirical literature. By drawing together recent empirical findings, and augmenting them with the results of new tests, answers to more general questions are suggested.

The paper is organized as follows. Section I contains a review of the theories designed to explain the relation between money supply announcements and asset prices. The review focuses on the role each of the theories assigns to expectations about the future supply of and demand for money, and the relation between these expectations, inflation forecasts, and the *ex ante* real rate. Section II presents empirical results on the reaction of asset prices to money supply announcements and compares them with the findings of previous studies. Results of tests based on the times-series behavior of expected money are also reported. In Section III, the competing theories are evaluated in light of the empirical evidence. Though the tests show that money stock announcements have an impact on the *ex ante* real rate, it is still not possible to conclude that monetary policy affects the *ex ante* real rate, because the announcements may function solely as signals which reveal information about other real variables.

## I. Money Supply Announcements and Asset Prices: The Theory

There are at least four channels through which information about the money stock can affect asset prices. To illustrate the linkages, it is useful to begin with the Fisher equation, (1), which states that the nominal interest rate equals the expected real rate plus expected inflation,

$$(1)\qquad i_t = E_t(r_{t+1}|I_t) + E_t(\pi_{t+1}|I_t),$$

where both expectations are conditional on the information available at time $t$. In terms of equation (1), money supply announcements affect interest rates by altering the information set and thereby changing expectations. It is important to stress that this is the *only* way that announcements can affect rates, since they have no impact on the actual money stock.

### A. *The Expected Inflation Hypothesis*

The most obvious link between money announcements and short-term interest rates is through expected inflation. (See Eugene Fama, 1975, for a detailed discussion.) The hypothesis is that announcement of an unanticipated jump in the money stock leads to expectations of higher inflation and thus to an increase in short-term interest rates, while announcement of an unanticipated drop in money has the reverse effect. One requirement of the expected inflation approach is that prices respond quickly to changes in money. If prices take several months to respond, as some authors have suggested, then rational agents would not alter their short-run inflation forecasts on the basis of a money supply announcement and rates of return on short-term assets like Treasury bills would not be affected.

The predicted response of long-term rates depends on how agents interpret the money supply announcements. If they permanently alter their anticipations regarding future money growth and future inflation, then long rates will change as much as short rates. Conversely, if long-term inflationary expectations remain largely unaffected by one week's report on the money stock, then the impact on long rates will be small. However, both short- and long-term rates should move in the same direction.

Turning to the foreign exchange market, a rise in the expected rate of inflation unambiguously implies that the dollar should depreciate against other currencies. *Ceteris paribus*, higher expected inflation reduces the demand for dollars; as a result, the dollar declines in the foreign exchange market.

The predicted response of stock prices is not so straightforward. If all markets were perfect and if there were no taxes, then changes in expected inflation would have little impact on stock prices because anticipated corporate cash flows, and the rate at which those cash flows are discounted, would be adjusted by offsetting amounts. Of course, this prediction breaks down if there are taxes (see Martin Feldstein, 1980), or if market imperfections induce firms to contract in nominal terms (see Kenneth French, Richard Ruback, and G. William Schwert, 1983). Furthermore, empirical studies by Charles Nelson (1976), Fama and Schwert (1977), and Schwert (1981), among others, show that stock returns are negatively correlated with both expected and unexpected inflation. However, Robert Geske and Richard Roll (1983) present evidence which indicates that the negative correlation arises because falling stock prices signal higher inflation, not because higher expected inflation leads to a drop in stock prices. To be conservative, therefore, the relation between stock prices and expected inflation is judged to be ambiguous. Fortunately, a definitive answer is not required for the empirical tests.

To summarize, the expected inflation hypothesis predicts that in response to an unexpected increase in the money supply: (*i*) short-term interest rates will rise; (*ii*) long-term rates will rise, but the increase will be less than for short-term rates unless the change in expected inflation is permanent; (*iii*) the dollar will depreciate against other major currencies; and (*iv*) stock prices may move up or down depending on the role of taxes, nominal contracting by firms and other markets imperfections.

### B. *The Keynesian Hypothesis*

A second possible link between money and asset prices is through the expected real rate. For this link to work, it must first be the case that *actual* innovations in money affect the *ex ante* real rate. The best known theory that posits such a link is the Keynesian liquidity preference model with sticky prices. The workings of the model are illustrated by assuming that the money demand function takes the simple form

$$M/P = f(Y, i), \tag{2}$$

where $M$ = the nominal stock of money, $P$ = the price level, and $Y$ = the level of real income.

A liquidity effect exists because the price level does not respond instantaneously to monetary shocks. For this reason, the interest rate must adjust to clear the money market. A sudden increase in the supply of nominal money balances, for instance, causes real money balances to rise because prices are sluggish. To clear the money market, the interest rate must fall to produce an offsetting increase in money demand. Since expected inflation will rise, if it is affected at all, the decline in the nominal rate must be due to a drop in the *ex ante* real rate. However, this liquidity effect is a short-run phenomenon. The interest rate starts to return to its initial level as soon as prices begin to adjust. When the adjustment of prices is complete, the interest rate is back at its starting point.

If actual changes in money affect the *ex ante* real rate, then announced changes will also have an impact if they alter anticipations regarding future Fed policy. The most common approach is to assume that the Fed is attempting to control the growth rate of monetary aggregates. In this context, an unexpected jump in the money stock leads market participants to believe that the Fed will have to tighten credit to offset the rise.

Because agents anticipate that future tightening will lead to higher rates via the liquidity effect, they bid aggressively for funds and drive up the current interest rate. Urich and Wachtel (1981) refer to this scenario as the policy anticipation effect.

If the policy anticipation effect is correct, money supply announcements should have a different impact on short-term interest rates compared to long-term interest rates. There are two possible cases. In the first instance, the anticipated period of monetary restraint is expected to be short-lived. In that case, only the spot rate and nearby forward rates are affected and the impact on thirty-year yields is minimal. On the other hand, if the anticipated period of monetary restraint is expected to be prolonged, distant forward rates, and hence long-term rates, will *fall*. The drop occurs because the liquidity effect is short-lived and because a prolonged period of monetary restraint will eventually reduce expected inflation. (I do not mean to rule out the possibility that expected inflation adjusts immediately, my point is simply to stress that economists of all persuasions would accept the proposition that prices and expectations respond in the long run.)

In the foreign exchange market, a rise in the *ex ante* real rate will cause the dollar to appreciate. As Rudiger Dornbusch (1976) and Engel and Frankel describe, a higher U.S. real rate induces an incipient capital inflow which causes the exchange rate to rise.

On the other hand, an increase in the *ex ante* real rate produced by monetary restraint should lead to lower stock prices for two reasons. First, the discount rate rises to reflect the higher real rate. Secondly, expected corporate cash flow will decline if agents believe that an increase in the real rate depresses economic activity.

In summary, the Keynesian hypothesis predicts that in response to announcement of an unexpected increase in the money supply: (*i*) short-term interest rates will rise; (*ii*) long-term rates will either decline or remain largely unaffected depending on whether the period of future monetary restraint is seen as being permanent or transitory; (*iii*) the dollar will appreciate against other major currencies; and (*iv*) stock prices will decline.

### C. *The Real Activity Hypothesis*

The next link between money supply announcements and asset prices is based on the hypothesis that the announcements provide information about future money *demand*.[1] For the link to operate money demand must depend on expected future output, as well as current output. This proposition is strongly supported by Fama's (1982) empirical work. Fama finds that when both $Y_t$ and $Y_{t+1}$ are included in the money demand function, $Y_{t+1}$ is always highly significant in the estimated equations. He argues that this result occurs because actual future output serves as a proxy for expected future output.

If money demand depends on expected future output, then individual agents cannot determine aggregate money demand, because other agents' expectations are unobservable. In this context, money supply announcements provide information about expected future output. For example, an unanticipated increase in the money stock tells agents that aggregate money demand is greater than they forecast. On the basis of this information, estimates of future real activity rise. With higher expected future output, the real rate must rise to clear the market for consumption and investment. In addition, the increase in expected output leads to a rise in the *current* demand for money. Assuming that prices adjust slowly and that the Fed does not accommodate the increased demand, the *ex ante* real rate must rise to clear the money market.

Vance Roley presents an example of the real activity hypothesis which is based on the current reserve settlement procedure. Under lagged reserve accounting, the new information provided by a larger than anticipated increase in money leads to an upward revision of expectations about the demand for reserves through Wednesday of the current week. With a fixed supply of nonborrowed reserves, short-term yields must rise to equilibrate the market for reserves. It should be stressed that Roley's example is a special case. The real activity hypothesis does not depend on lagged reserve accounting. In fact, *the real activity hypothesis can hold even if money has no affect on the real rate* as long as money stock announcements provide information about expected future output.

---

[1] Nichols, Small, and Webster develop a variant of this approach. Similar models have also been analyzed by Gikas Hardouvelis (1982) and Vance Roley (1982).

According to the real activity hypothesis, the reaction of long-term rates to money supply announcements depends on how the announcements affect expected output over the life of the bond. An expected spike in output, for instance, would have little impact on the long-term real rate. In general, unless agents interpret the information in the money supply announcement to imply that the shock to output is permanent, long rates will move in the same direction as short rates, but to a lesser degree.

An expected rise in real activity will cause the dollar to appreciate for two related reasons. First, the increase in the demand for dollars directly affects the exchange rate. Second, the higher *ex ante* real rate associated with greater money demand leads to an incipient capital inflow.

Information which causes agents to revise upwards their forecasts of future real activity should cause stock prices to rise. Though the discount rate will rise to reflect the higher *ex ante* real rate, the increase is more than offset by the growth of expected corporate cash flow. This must be the case because the reason for the higher real rate is an expected increase in output. In an empirical study (1981), furthermore, Fama finds a highly significant relation between rates of return on common stock and innovations in real activity.

To conclude, the real activity hypothesis predicts that in response to an unexpected increase in the money supply: (*i*) short-term interest rates will rise; (*ii*) long-term rates will rise, but the increase will be less than for short-term rates unless the change in expected output is permanent; (*iii*) the dollar will appreciate against other major currencies; and (*iv*) stock prices will rise.

### D. *The Risk Premium Hypothesis*

In his classic paper (1958), James Tobin argues that money demand depends on agents' risk aversion and on their perception of the riskiness of competing assets. Incorporating Tobin's insight, the money demand function can be written as

$$(3)\qquad M/P = f(i, Y, \theta),\ df/d\theta > 0,$$

where $\theta$ is a parameter which depends on both aggregate risk aversion and aggregate beliefs about future security returns. The parameter $\theta$ is defined so that an increase in risk aversion or an increase in the risk of financial assets causes $\theta$ to rise. Assuming that each agent knows only his own risk preferences and beliefs, money supply announcements provide information about other agents' risk preferences and beliefs.[2] For example, holding real income constant, announcement of an unexpected jump in the money stock reveals that aggregate risk aversion has risen or that competing assets are now perceived to be riskier. In either case, the information causes required real returns to rise and asset prices to fall. In addition, the increase in money demand associated with a jump in $\theta$ puts upward pressure on the real rate.

A scenario consistent with the risk-aversion hypothesis is that large positive innovations in money put pressure on the Fed to "do something," and this makes future policy more uncertain. Uncertainty about monetary policy, in turn, increases inflation risk which causes risk premiums to rise. The impact of money stock announcements on long-term bond yields is large because of the degree to which these securities are exposed to inflation risk. Conversely, a negative innovation in money gives the Fed added leeway and reassures the market that no fundamental change in policy is forthcoming.

One weakness of the risk premium approach is that it does not adequately distinguish between "money" and short-term interest bearing assets such as Fed funds and overnight repos. It is possible that an increase in $\theta$ will also cause the demand for these assets to rise relative to the demand for stock and long-term bonds. If there is a net

---

[2] I am indebted to Richard Roll for suggesting the risk premium hypothesis as a possible alternative.

TABLE 1—PREDICTED RESPONSES TO ANNOUNCEMENT OF AN UNEXPECTED INCREASE IN $M1$

| | Expected Inflation Hypothesis | Keynesian Hypothesis | Real Activity Hypothesis | Risk-Aversion Hypothesis |
|---|---|---|---|---|
| A. Predicted Response of Asset Prices | | | | |
| Three-month $T$-bill *yield* | Rise | Rise | Rise | Ambiguous |
| Thirty-year $T$-bond *yield* | Rise | Ambiguous | Rise | Rise |
| S&P 500 Stock Index | Ambiguous | Drop | Rise | Drop |
| German mark (dollar price) | Rise | Drop | Drop | No change |
| B. Predicted Response of Expected Money | Above average increase | Decrease | No prediction | No prediction |

increase in the demand for short-term assets, then their yields will fall. In general, therefore, the relation between $\theta$ and the demand for interest bearing assets will be a function of the security's maturity. Though it is clear that the yields on thirty-year bonds will increase as $\theta$ grows, there may be some maturity at which a switch occurs and the yield becomes a decreasing function of $\theta$. This is an empirical question.

If the information contained in money supply announcements causes risk premiums to increase, it is clear that stock prices will fall. In the case of foreign exchange, however, the situation is less obvious. Though the *ex ante* real rate rises, this will not cause an incipient capital inflow if the higher real rate is fair compensation for the greater risk of holding dollar assets. To summarize, therefore, the risk-aversion hypothesis predicts that in response to announcement of an unexpected increase in money: (*i*) short-term interest rates may rise or fall. However, the greater the maturity of the security, the more likely it is that its yield will increase; (*ii*) long-term rates will rise; (*iii*) stock prices will fall; and (*iv*) the dollar will remain unaffected.

### E. *The Role of Expected Money*

In some cases, the competing hypotheses also make different predictions regarding the reaction of expected money to announced innovations. According to the expected inflation hypothesis, announcement of unanticipated increases in $M1$ lead agents to forecast faster money growth and higher future inflation. Conversely, the Keynesian hypothesis predicts that an unanticipated jump in the money stock causes agents to anticipate tightening by the Fed. Thus these two theories can be distinguished by examining the cross-correlation function between announced innovations and changes in expected money. If the expected inflation hypothesis is correct, the correlation between current innovations and future expected money will be positive, while the Keynesian hypothesis predicts that the correlation will be negative. The real activity and risk premium hypotheses make no explicit predictions regarding the response of expected money to announced innovations.

### F. *Predictions of the Theories*

The predictions of the competing hypotheses are summarized in Table 1. Notice that despite the ambiguities, no two hypotheses predict the same response of asset prices to money supply announcements. Thus, by studying the *simultaneous* reactions of all four assets to the announcements, at least three of the hypotheses can be rejected. The results can then be confirmed by checking whether they are consistent with the behavior of expected money.

TABLE 2—SUMMARY STATISTICS[a]

| | Mean | Standard Deviation | $\rho_1$ | $\rho_2$ | $\rho_3$ | $\rho_4$ |
|---|---|---|---|---|---|---|
| Sample period: January 5, 1978–October 4, 1979 | | | | | | |
| *EM* | 0.12 | 0.45 | 0.00 | −0.10 | 0.12 | 0.17 |
| *UM* | 0.20 | 0.35 | −0.10 | −0.21 | −0.02 | 0.02 |
| *DTB* | 2.58 | 10.61 | 0.11 | −0.13 | 0.07 | 0.05 |
| *DLTB* | 0.69 | 2.71 | 0.05 | −0.10 | −0.21 | −0.14 |
| *DSP* | 0.14 | 0.66 | 0.03 | 0.03 | −0.02 | 0.03 |
| *DDM* | 0.05 | 0.55 | 0.16 | 0.00 | −0.15 | −0.03 |
| Sample period: October 11, 1979–December 18, 1981 | | | | | | |
| *EM* | 0.06 | 0.26 | −0.03 | 0.03 | 0.05 | 0.16 |
| *UM* | −0.04 | 0.58 | −0.09 | 0.03 | −0.03 | 0.21 |
| *DTB* | 8.68 | 40.80 | 0.06 | 0.17 | 0.11 | 0.17 |
| *DLTB* | 4.57 | 18.60 | 0.08 | 0.00 | −0.04 | 0.08 |
| *DSP* | −0.14 | 1.13 | 0.11 | −0.06 | 0.06 | −0.01 |
| *DDM* | 0.04 | 0.91 | 0.09 | 0.04 | 0.03 | −0.03 |

*Notes: UM* = The unexpected change in the money supply in percent; *EM* = the expected change in the money supply in percent; *DTB* = the change in the yield on the latest three-month Treasury bill, in basis points; *DLTB* = the change in the thirty-year bond yield, in basis points; *DSP* = the percentage change in the S&P 500 stock price index; *DDM* = the percentage change in the dollar price of the German mark; *DG* = the percentage change in the price of gold; $\rho_i$ = sample autocorrelation for lags $i = 1,2,3,4$.

[a] All changes in asset prices are measured from the close before the money supply announcement to the close after the announcement.

## II. Empirical Results

### A. *Asset Prices*

This section draws on both original research and the existing literature. The goal is to bring together all the evidence related to the predictions in Table 1. Though no previous paper has examined the simultaneous response of all four assets to money supply announcements, results exist for each of the individual assets over a variety of sample periods. Therefore, my findings can be confirmed by checking them against the results of related studies.

The predictions in Table 1 are tested by regressing changes in the prices of three-month Treasury bills, thirty-year Treasury bonds, German marks, and the S&P 500 stock index on both the expected and unexpected components of the money supply announcements. Changes in asset prices are measured from the market close before the money supply announcement to the close after the announcement. The Treasury bill and Treasury bond data were provided by the Federal Reserve Bank of San Francisco. The stock price data are taken from the *Daily Stock Price Record* and the foreign exchange data are taken from the *International Monetary Market Yearbook*. The expected change in the money supply is assumed to be equal to the median of the Money Market Services survey for *M*1 (previously *M*1*B* and before that *M*1). The Money Market Services survey is described in detail by Urich and Wachtel (1982b). Previous studies indicate that the survey median is a reasonable proxy for expectations. My (1983) article finds that the correlation between survey forecast errors and changes in asset prices is higher than the correlation between changes in asset prices and forecast errors derived from an ARIMA model. In addition, Engel and Frankel, Jacob Grossman (1981), Roley, Douglas Pearce and Roley (1982), Urich (1982), and Urich and Wachtel (1982a) all report that forecast errors derived from the survey data are highly correlated with changes in the prices of various assets.

The sample runs from January 5, 1978 to December 18, 1981. This four-year period is divided into two intervals to take account of the Federal Reserve's stated change in operating procedure on October 6, 1979.

TABLE 3—MONEY SUPPLY ANNOUNCEMENTS AND CHANGES IN ASSET PRICES[a]

$DA_t = a_0 + a_1UM_t + a_2EM_t + u_t$

| $DA$ | $a_0$ | $a_1$ | $a_2$ | $R^2$ |
|---|---|---|---|---|
| Sample period: January 5, 1978–October 4, 1979 | | | | |
| *DTB* | 2.30 | 1.92 | 0.24 | .063 |
| | (1.72) | (0.77) | (0.07) | |
| *DLTB* | 0.81 | −0.36 | −0.42 | .066 |
| | (2.38) | (−0.56) | (−0.52) | |
| *DSP* | 0.01 | −0.18 | 0.13 | .063 |
| | (1.62) | (−1.19) | (0.66) | |
| *DDM* | −0.001 | 0.10 | −0.06 | .016 |
| | (−0.72) | (0.74) | (−0.35) | |
| Sample period: October 11, 1979–December 18, 1981 | | | | |
| *DTB* | 7.21 | 30.46 | −5.36 | .234 |
| | (1.97) | (4.33) | (−0.35) | |
| *DLTB* | 3.48 | 15.01 | 3.16 | .242 |
| | (2.09) | (4.68) | (0.45) | |
| *DSP* | −0.001 | −0.55 | −0.25 | .080 |
| | (−0.84) | (−2.63) | (−0.55) | |
| *DDM* | 0.001 | −0.40 | −0.42 | .050 |
| | (0.16) | (−2.35) | (−1.11) | |

*Notes:* See Table 2.

[a]$t$-statistics are shown in parentheses.

Summary statistics for the variables are presented in Table 2. The table dramatically illustrates the increased volatility of interest rates after October 6, 1979. For instance, the standard deviation of the change in the long-term bond rate on the day $M1$ is announced increased by 6.7 times. On the other hand, the standard deviation of unanticipated changes in money increased only 1.7 times and the standard deviation of expected money actually declined.

The regression results are reported in Table 3. It is convenient to summarize the results as a series of "empirical propositions." This makes for easy comparison with other studies and facilitates examination of the competing theories in the next section.

PROPOSITION 1: *The asset markets are efficient with respect to money supply announcements in the sense that only the unexpected component of the announcement is correlated with price changes.*

The proposition is clearly supported by Table 3. The coefficient of the expected change in the money supply is not significant for any of the assets. Similar findings are reported by Grossman, Pearce and Roley, Roley, Urich, and Urich and Wachtel (1981, 1982a) for individual assets. In some cases residuals from ARIMA models, rather than survey forecast errors, were used as the right-hand variable, marginally reducing the correlations, but not affecting the conclusions.

PROPOSITION 2: *After October 6, 1979, there is a highly significant positive correlation between changes in short-term interest rates and the unexpected component of money supply announcements. Prior to that date, however, the relation is weak.*

The results for three-month Treasury bills in Table 3 show that the $t$-statistic on $UM$ is 0.77 before October 6 and 4.33 after that date. Roley also uses three-month bill yields, but measures the change in the interest rate from 3:30 P.M. (EST) to 5:00 P.M., rather than from close to close. (The money supply announcement is at 4:10 P.M.) He finds that the $t$-statistic for the coefficient of $UM$ rises from 2.44 before October 6 to 6.84 after October 6. The effect is not unique to three-month Treasury bills. Studies by myself

(1983) and Hardouvelis show that yields on securities from overnight Fed funds to one-year Treasury bills are positively correlated with announced innovations in the money stock. Further supporting evidence is provided by Urich, and by Urich and Wachtel (1982b). The only debatable issue in the literature is whether unexpected money supply announcements had a significant impact on short-term interest rates before October 6, 1979. Urich and Wachtel (1981) and Roley, for instance, find some marginally significant coefficients. More importantly, both Roley and Urich and Wachtel (1982a) find that the correlations are higher when the unanticipated money supply announcement pushes money growth rates outside the targets set by the Federal Reserve. Finally, these results hold for both survey forecast errors and ARIMA model residuals.

When the money supply figures are released, the Fed also announces any revision in the previous week's estimate. It is possible, therefore, that unexpected revisions account for part of the movement in asset prices. The only test of this hypothesis is conducted by Roley. Assuming that the expected revision is zero, Roley finds that unanticipated revisions have no significant impact on Treasury bill yields. In the remainder of this paper revisions are ignored.

PROPOSITION 3: *The reaction of long-term bond yields is similar to the reaction of short-term bill yields. Though the regression coefficient is smaller, the t-statistic is just as high. Once again, the effect is only observed after October 6, 1979.*

If long-term yields respond to money supply announcements it must be the case that forward rates, as well as current spot rates, are affected. Hardouvelis directly tests this hypothesis. In light of Table 3, it is not surprising that he finds that money supply announcements have a significant impact on the ten-year forward rate, ten years forward. My (1983) article studied the reaction of a variety of Treasury securities to money supply announcements. I found that the correlation between yield changes and announced innovations in money is essentially constant for securities with maturities ranging from one day to thirty years.

PROPOSITION 4: *Announced innovations in the money supply are negatively correlated with changes in the dollar price of foreign exchange, but only after October 6, 1979.*

The results for the German mark presented in Table 3 are similar to the findings of Engel and Frankel. My (1982) article and Hardouvelis report that the correlations are also negative for the British pound, Canadian dollar, French franc, Japanese yen, and Swiss franc, though the estimates are significant only in the case of the yen and Swiss franc.

PROPOSITION 5: *After October 6, 1979, announced innovations in the money supply are negatively correlated with stock price returns. There is also marginal evidence of negative correlation before October 6.*

Though the results in Table 3 for the S&P 500 index show that announced money supply innovations are negatively correlated with stock prices both before and after October 6, 1979, the correlation is not significant before the Fed's change in operating procedure. Using the Dow Jones Industrial Average and a sample period from September 1977 to October 1981, Pearce and Roley report a $t$-statistic of $-3.6$ for the coefficient of $UM$ before October 6 and a $t$-statistic of $-5.2$ after that date.

PROPOSITION 6: *Propositions 1 through 5 fail to hold for the monetary base. Announced innovations in the base have no significant impact on any of the asset prices either before or after October 6, 1979.*

Results for the monetary base are reported in Table 4. Since survey data are not available for the base, residuals from an ARIMA model are used instead.[3] As noted earlier, this substitution cannot explain the dif-

---

[3] A separate ARIMA model is estimated for each period. The data are weekly observations on the percentage change in the base, $DMB$. The entire sample is used in the estimation, so the models incorporate infor-

TABLE 4—MONETARY BASE ANNOUNCEMENTS AND CHANGES IN ASSET PRICES[a]

$DA_t = a_0 + a_1 UB_t + a_2 EB_t + u_t$

| $DA$ | $a_0$ | $a_1$ | $a_2$ | $R^2$ |
|---|---|---|---|---|
| Sample period: January 5, 1978–October 4, 1979 | | | | |
| *DTB* | 2.89 | −0.91 | 1.78 | .067 |
| | (2.30) | (−0.25) | (0.32) | |
| *DLTB* | 0.44 | 1.63 | 1.84 | .133 |
| | (1.40) | (1.79) | (1.31) | |
| *DSP* | 0.002 | −0.10 | −0.44 | .078 |
| | (2.39) | (−0.46) | (−1.29) | |
| *DDM* | −0.001 | 0.14 | 0.16 | .025 |
| | (−1.39) | (0.76) | (0.55) | |
| Sample period: October 11, 1979–December 18, 1981 | | | | |
| *DTB* | 6.11 | 8.07 | −51.24 | .087 |
| | (1.50) | (0.88) | (−2.10) | |
| *DLTB* | 2.64 | 8.65 | −20.23 | .114 |
| | (1.44) | (2.10) | (−1.84) | |
| *DSP* | −0.002 | 0.15 | −0.64 | .022 |
| | (−1.39) | (0.06) | (−0.94) | |
| *DDM* | 0.000 | −0.31 | 0.34 | .024 |
| | (0.20) | (−1.50) | (0.62) | |

*Notes:* See Table 2; *UB* = the unexpected change in the monetary base in percent, from an ARIMA model; *EB* = the expected change in the money supply in percent, from an ARIMA model.

[a]Standard errors are in parentheses.

TABLE 5—CROSS CORRELATIONS BETWEEN ANNOUNCED INNOVATIONS AND EXPECTED MONEY[a]

(Correlation between $UM(t)$ and $EM(t+j)$ for $j = -3$ to $+4$)

| $\rho_{-3}$ | $\rho_{-2}$ | $\rho_{-1}$ | $\rho_0$ | $\rho_1$ | $\rho_2$ | $\rho_3$ | $\rho_4$ |
|---|---|---|---|---|---|---|---|
| Sample period: January 5, 1978–October 4, 1979 | | | | | | | |
| −0.14 | −0.04 | 0.04 | 0.07 | −0.53 | −0.10 | −0.03 | 0.03 |
| (0.11) | (0.11) | (0.10) | (0.10) | (0.10) | (0.11) | (0.11) | (0.11) |
| Sample period: October 11, 1979–December 18, 1981 | | | | | | | |
| −0.20 | −0.14 | −0.11 | 0.15 | −0.52 | −0.15 | −0.11 | 0.09 |
| (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) |

*Notes:* $DM(t)$ = the percentage change in $M1$ in period $t = EM(t) + UM(t)$.

[a]Standard errors are in parentheses.

ference between Tables 3 and 4. When ARIMA models are used for $M1$, the results are almost identical to those reported in Table 3.

### B. *Expected Money*

Cross correlations between announced innovations and expected changes in money are presented in Table 5. The correlation between $UM(t)$ and $EM(t+1)$ is −0.52 which demonstrates that following an innovation, agents expect an offsetting change the next week. Though only eight cross correlations are reported, the function was computed out to a $t+36$. Except for the spike at $t+1$, the cross-correlation function is essentially flat. The second largest correlation is 0.23 at week $t+15$. In all only 4 of the 40 estimated correlations are significant at the 5 percent level.

Surprisingly, there is no evidence of a shift on October 6, 1979. The cross-correlation

---

mation unavailable to investors. The two models are

Before October 6, 1979: $DMB_t = 0.161 + u_t + 0.37u_t - 0.45u_{t-1}$;

After October 6, 1979: $DMB_t = -0.34DMB_{t-1} + 0.135 + u_t$.

function is almost identical for the two periods. Finally, the contemporaneous correlation between expected money and the survey forecast error is insignificantly different from zero in both periods, which supports the view that the survey data are rational forecasts.

### III. Comparing the Theories

Comparing the predictions in Table 1 with the empirical results reveals that none of the theories is completely satisfactory. The expected inflation hypothesis fails to explain why the dollar appreciates after announcement of an unanticipated increase in $M1$, and does not adequately account for the simultaneous decline in stock prices. The Keynesian approach is inconsistent with the high positive correlation between changes in long-term bond yields (including distant forward rates) and announced innovations in the money supply. The real activity hypothesis cannot explain the decline in the stock market associated with an unexpected jump in the money supply. Finally, the risk premium hypothesis does not account for the appreciation of the dollar. In addition, the significant positive correlation between changes in the Fed funds rate and announced innovations in money contradicts one interpretation of the risk premium hypothesis.

The monetary base results indicate that the market is reacting to information about money demand, not supply. Because the monetary base and its components are revealed a week before the $M1$ figures are released, the $M1$ announcement provides no new data on Fed policy. Nonetheless, the market responds to announced innovations in $M1$, but not the base. This means that if the new information is related to the supply of money, it must be information about supply shocks originating within the banking system.

The base results cast doubt on the expected inflation hypothesis. According to the hypothesis, money supply announcements move interest rates because they affect both short- and long-run expectations about money growth. This implies that announced innovations in the base should be more highly correlated with asset prices than announced innovations in $M1$, because the base announcements provide new information about Fed policy. However, the empirical findings reveal that the reverse is true.

The dramatic shift in the market response to money supply announcements after October 6, is difficult to reconcile with the expected inflation hypothesis. If the money supply announcements are providing information about future money growth, there is no obvious reason why the Fed's stated intention to control monetary aggregates should induce a positive correlation between announced innovations in money and changes in interest rates. In fact, it is more reasonable to conclude that the correlation would decline because week-to-week variation in the aggregates would no longer provide information about long-run policy.

If the money supply announcements are providing agents with information about future real activity, it is hard to see why a sudden shift should occur on October 6, 1979. There is no reason to believe that the money demand function shifted, or that real shocks intensified, on that date. Thus it is unclear why the $M1$ announcement suddenly becomes a better signal of future real activity. The hypothesis can be salvaged, however, by combining it with the Keynesian model. Since an expected increase in future activity leads to a rise in *current* money demand, the real rate must rise to clear money unless the Fed intervenes. Before October 6, the real rate does not rise because agents expect the Fed to accommodate demand shocks. After October 6, however, accommodation is no longer anticipated. It is important to note that this interpretation only works if money has an impact on the real rate as posited by the Keynesian model.

The risk premium hypothesis has a similar problem. What happened on October 6 to make risk premiums more sensitive to money supply announcements? One explanation is that the dramatic change in policy on that date informed agents that the rules for conducting monetary policy are variable. As a result, investors begin studying weekly announcements for signs of further changes in the rules. This increases the sensitivity of the premium, particularly the premium on long-term fixed rate securities, to money supply announcements.

Finally, the shift on October 6 is predicted by the Keynesian model. According to the Keynesian theory, the policy anticipation effect only operates when the Federal Reserve is attempting to control the growth rate of monetary aggregates. Prior to October 6 the announcement effect is weak because the Fed was focusing on the Fed funds rate, not the money stock.

Similarly, the Keynesian model explains why the market reaction is larger when an announcement pushes money growth rates outside the limits set by the Fed. When the limits are breached, agents expect the Fed to respond more aggressively to offset money surprises. On the other hand, for the risk premium hypothesis to explain the "target effect," it must be postulated that policy becomes less predictable when growth rates rise above the upper limit, but that policy becomes more predictable when growth rates fall below the lower limit. If the expected inflation hypothesis holds, it must be the case that when money growth rates rise above the upper limit, agents expect even *faster* growth in the future. Neither of these two explanations is very plausible. The real activity hypothesis can be reconciled with the target effect by postulating that the Fed partially accommodates money demand shocks, but only when money growth rates are within the set limits.

The behavior of expected money offers apparent support for the Keynesian hypothesis. The cross-correlation function shows that agents expect large innovations in money to be quickly offset. However, the only significant negative correlation is between an announced innovation in week $t$, and the expected change in money in week $t+1$. Beyond week $t+1$, the cross-correlation function is flat. This pattern of correlations is more easily explained by random week-to-week fluctuations in variables like float and Treasury deposits, than by the Keynesian hypothesis. For instance, suppose transportation problems produce an unexpected increase in float in week $t$. Because such difficulties are generally overcome in a week, rational agents forecast a drop in $M1$ in week $t+1$. The fact that there is no shift in the cross-correlation function on October 6 is consistent with this explanation.

The tests thus far have been based on the assumption that only one of the competing hypotheses holds. Since the theories are not mutually exclusive, however, it is possible that some combination of them more accurately characterizes reality. For example, Hardouvelis argues that both the Keynesian hypothesis and the expected inflation hypothesis are true, but that the liquidity effect dominates the short-run market reaction and the inflation premium dominates the long-run market reaction. In addition to the fact that it is intuitively difficult to understand how the same announcement leads agents to expect monetary restriction in the short run, but monetary ease in the long run, the Hardouvelis argument has two technical problems. First, the cross-correlation function between announced innovations and expected money is inconsistent with his interpretation. Second, the monetary base results do not support the view that announcement effect is based on new information about the Fed's discretionary policy.

Because it is quite possible that all four of the hypotheses are true to some degree, there is no reason to incorporate only two into an eclectic theory. There are a wide variety of combinations of the hypotheses that are consistent with the empirical results reported here. Without designing new tests, however, there is no way to distinguish between competing eclectic theories.

Since the expected inflation hypothesis alone fails to explain the findings, it is tempting to assert that the results show that monetary shocks affect the *ex ante* real rate, but this conclusion does not follow. It is possible that the real rate moves solely because the announcements convey information about real variables such as expected output and risk preferences. The situation is analagous to the dividend announcement problem in finance. It is well known that announced innovations in dividends are positively correlated with stock prices, but it is unclear whether changes in dividends move stock prices, whether the dividend announcements convey information about other real variables like future earnings, or whether a combination of the two is true.

## IV. Summary and Conclusions

Studying the reaction of asset prices to money supply announcements presents a unique opportunity for testing the hypothesis that monetary policy affects the *ex ante* real rate. First, using the announcements removes ambiguity about the direction of causation, since the money supply announcements, as opposed to actual changes in the money stock, are exogenous with respect to prices and interest rates. Second, use of the announcements makes it possible to reduce the observation interval to one trading day or less. This greatly increases the power of statistical tests.

There are at least four theories which explain why money supply announcements move asset prices. First, the expected inflation hypothesis states that the announcements alter agents' inflation forecasts. Second, the Keynesian hypothesis predicts that in response to an announced innovation in the money stock, agents expect the Fed to take offsetting action. The anticipated policy response, in turn, causes the real rate to change in the direction predicted by the liquidity preference model. For example, an expected tightening of credit leads to an increase in the *ex ante* real rate. Third, the real activity hypothesis alleges that money supply announcements provide the market with information about future output, and thereby future money demand. An unexpected jump in the money stock, for instance, signals higher output and greater money demand. This information produces an increase in the *ex ante* real rate. Finally, the risk premium hypothesis states that money stock announcements alter the required real return on financial assets by providing the market with information about aggregate risk preferences and beliefs.

Two different tests of the hypotheses are conducted. The first is based on the simultaneous response of the prices of short-term bills, long-term bonds, common stock, and foreign exchange to announcements of both $M1$ and the monetary base. None of the hypotheses explains the reaction of all four assets. The second test, based on the relation between announced innovations in money and expected money, is consistent with the Keynesian hypothesis, but the results can be more readily explained by random variation in float and Treasury deposits.

The tests show that money supply announcements have an impact on the real rate, but they do not allow us to conclude that monetary shocks affect the real rate. This apparent paradox arises because the announcements also function as signals which reveal information about real variables such as expected future output and risk preferences. Further tests, using data in addition to money supply announcements, are required to separate those hypotheses that rely on the signaling effect from those which assume that money affects the real rate.

## REFERENCES

**Cornell, Bradford,** "Money Supply Announcements, Interest Rates, and Foreign Exchange," *Journal of International Money and Finance*, August 1982, *1*, 201–08.

______, "Money Supply Announcements and Interest Rates: Another View," *Journal of Business*, January 1983, *56*, 1–24.

**Dornbusch, Rudiger,** "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–76.

**Dwyer, Gerald P.,** "Is Variation of the Expected Real Interest Rate Unpredictable?," *Journal of Monetary Economics*, January 1981, *8*, 59–84.

**Engel, Charles and Frankel, Jeffrey,** "Why Money Announcements Move Interest Rates: An Answer From the Foreign Exchange Market," unpublished working paper, University of California-Berkeley, 1982.

**Fama, Eugene F.,** "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

______, "Stock Returns, Real Activity, Inflation and Money," *American Economic Review*, September 1981, *71*, 545–65.

______, "Inflation, Output and Money," *Journal of Business*, April 1982, *55*, 201–32.

______ **and Gibbons, Michael R.**, "Inflation, Real Returns and Capital Investment," *Journal of Monetary Economics*, May 1982, *9*, 297–324.

______ **and Schwert, G. William**, "Asset Returns and Inflation," *Journal of Financial Economics*, November 1977, *5*, 115–46.

**Feldstein, Martin**, "Inflation and the Stock Market," *American Economic Review*, December 1980, *70*, 839–47.

**French, Kenneth R., Ruback, Richard S. and Schwert, G. William**, "Effects of Nominal Contracting on Stock Returns," *Journal of Political Economy*, February 1983, *91*, 70–96.

**Geske, Robert, and Roll, Richard**, "The Fiscal and Monetary Linkage Between Stock Returns and Inflation," *Journal of Finance*, March 1983, *38*, 1–33.

**Grossman, Jacob**, "The Rationality of Money Supply Expectations and the Short-Run Response of Interest Rates to Monetary Surprises," *Journal of Money, Credit, and Banking*, November 1981, *13*, 409–24.

**Hardouvelis, Gikas A.**, "Market Perceptions of Federal Reserve Policy and the Weekly Monetary Announcements," unpublished working paper, University of California-Berkeley, 1982.

**Mishkin, Frederic S.**, (1981a) "Monetary Policy and Long-Term Interest Rates: An Efficient Markets Approach," *Journal of Monetary Economics*, January 1981, *7*, 29–55.

______, (1981b) "The Real Interest Rate: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy: The Costs and Consequences of Inflation*, Spring 1981, *15*, 153–202.

______, "Monetary Policy and Short-term Interest Rates: An Efficient Markets Rational Expectations Approach," *Journal of Finance*, March 1982, *37*, 63–72.

**Nelson, Charles R.**, "Inflation and Rates of Return on Common Stocks," *Journal of Finance*, May 1976, *31*, 471–83.

**Nichols, Donald A., Small, David H., and Webster, Charles E., Jr.**, "Why Interest Rates Rise when an Unexpectedly Large Money Stock is Announced," *American Economic Review*, June 1983, *73*, 383–88.

**Pearce, Douglas P. and Roley, Vance V.**, "The Reaction of Stock Prices to Unanticipated Changes in Money," unpublished working paper, Federal Reserve Bank of Kansas City, 1982.

**Roley, Vance V.**, "The Response of Short-term Interest Rates to Weekly Money Announcements," unpublished working paper, Federal Reserve Bank of Kansas City, 1982.

**Schwert, G. William**, "The Adjustment of Stock Prices to New Information About Inflation," *Journal of Finance*, March 1981, *36*, 15–30.

**Tobin, James**, "Liquidity Preference as Behavior Towards Risk," *Review of Economics and Statistics*, February 1958, *25*, 65–86.

**Urich, Thomas J.**, "The Information Content of Weekly Money Supply Announcements," *Journal of Monetary Economics*, July 1982, *10*, 73–89.

______ **and Wachtel, Paul**, "Market Response to Weekly Money Supply Announcements in the 1970s," *Journal of Finance*, December 1981, *36*, 1063–72.

______ **and** ______, (1982a) "The Effects of Inflation and Money Supply Announcements on Interest Rates," unpublished working paper, New York University, 1982.

______ **and** ______, (1982b) "The Structure of Expectations of the Weekly Money Supply Announcement," unpublished working paper, New York University, 1982.

**Chicago Mercantile Exchange**, *International Monetary Market Yearbook*, Chicago, various years.

**New York Stock Exchange**, *Daily Stock Price Record*, New York, various issues.

# Implicit Contracts in the Absence of Enforcement and Risk Aversion

By CLIVE BULL*

Implicit contract theory (Martin Neil Baily, 1974; Costas Azariadis, 1975) grew out of a belief that the traditional neoclassical view of the labor market as simply a spot auction market could not explain such observed market phenomena as sticky wages and wages in excess of marginal revenue product. The great insight of these authors was to see that the absence of a complete set of contingent claims markets, in particular, for labor services and human capital, presented firms and workers with a mutually profitable opportunity to join in long-term, state-contingent, labor supply contracts which would replace spot auction markets for labor. Further analysis of these contracts has shown that, at least when information is asymmetric between the contracting parties, the optimal contracts struck may result in some of the labor market behavior that so puzzled economists who viewed it through spot auction market glasses (Sanford Grossman and Oliver Hart, 1981; Bengt Holmstrom, 1983; Jerry Green and Charles Kahn, 1983).

While the successes of this line of inquiry are hard to overstate, there are still some criticisms which can be leveled at it; the most serious of which will be dealt with here. As Hart has pointed out, only a limited proportion of all labor transactions are carried out under any type of formal, multiperiod contract and, moreover, few of those reach the complexity of detail which the contract literature suggests is optimal. Indeed, the whole thrust of the term "implicit contract" is to suggest that although a formal contract may not exist, or if it exists, it may be too simple, the two sides of the trade are nevertheless acting "as if" one existed, that is, we explain their otherwise puzzling behavior by the imposition of an extra, unobservable constraint subject to which they maximize. However, explicit contracts constrain behavior only if they can be enforced, at least probabilistically. This immediately raises the major issue of whether implicit contracts, which are by definition unenforceable, can affect economic behavior at all. So, to quote Hart "...if one appeals to the idea of an implicit contract, one has to say how it is enforced" (1983, p. 23).

This question of the enforceability of implicit contracts has recently received considerable attention from Holmstrom (1981, 1983). Prior to Holmstrom's work, it had been assumed that breach of the implicit contract could be prevented only by assuming that transaction costs between each of the implicitly contracting parties and the labor market were so high as to prevent them trading with the market rather than between themselves, thereby removing any incentive to breach the contract.[1] Holmstrom has demonstrated very elegantly that this is true only if the contract extends for one period. In a multiperiod contract, one party can fully insure the other party against the possibility that the first party will breach the contract at a future date. This is achieved by front-end loading. As an example, labor can insure an employer against the possibility that it will quit in the second period rather than fulfill the implicit contract by accepting a wage in the first period below the market wage. Moreover, Holmstrom (1981) has shown that, even after payment of this insurance pre-

---

*Department of Economics, New York University, 269 Mercer Street, 7th Floor, New York, NY 10003. I thank Costas Azariadis, Jim Brown, Janusz Ordover, Sherwin Rosen, Andrew Schotter, and Bruno Stein for comments and criticisms. I owe a special debt to Jess Benhabib and Roman Frydman for their suggestions and perseverance. I retain all responsibility for any errors.

[1] This was the approach used by Baily, though he imposed transactions costs only on the part of labor which leaves the firm's incentive to breach undiminished.

mium, there are gains from trade for both parties to the implicit contract and so it will dominate spot market trades.[2]

Although front-end loading is an elegant and practical way to attack the enforcement problem, it is, unfortunately, inherently one-sided. Through front-end loading, either the workers can indemnify the employer against future breach of the contract on their part, or the employer can indemnify the workers against future breach of the contract on his part, but both cannot do this simultaneously. Thus, by itself, recognizing the multiperiod nature of the contracts and appealing to front-end loading will not solve the enforcement problem.

As a complement to the use of front-end loading to indemnify the employer, several authors have suggested that concern by the firm for its reputation or brand name in the labor market may lead the employer to fulfill his part of the contract (Hart; Holmstrom, 1981, 1983; Azariadis and Joseph Stiglitz, 1983). In other words, an appeal is made to a third party, here the market rather than the court, for enforcement of the implicit contract. Although, in principle, reputation could induce contract-consistent behavior, the conditions under which this will occur seem to be too stringent for this to be a significant means of enforcement, except in very specialized labor markets.[3]

If one cannot rely on brand name or other types of third party enforcement of implicit contracts, nor on front-end loading, it appears that one must rely on transactions costs on both sides of the contract to prevent breaches and so allow implicit contracts to affect behavior, that is, for implicit contracts to be a viable alternative to spot market trade. As will be shown below, this conclusion is wrong, provided we realize that certain relevant markets for current services are missing.

One of the central ideas in the internal labor market literature in particular, and in the industrial relations literature in general, is that labor markets are incomplete (Peter Doeringer and Michael Piore, 1971; Oliver Williamson, Michael Wachter, and Jeffrey Harris, 1975). The firm purchases a bundle of labor services when it hires a worker and simultaneously offers the worker a bundle of goods and services as a compensation package. Moreover, for technological reasons many of the components of these bundles cannot be split off and traded separately. For instance, labor services such as work effort, initiative, and honesty cannot be purchased separately from each other, but only tied in to the purchase of a certain person's labor time. Similarly, the worker cannot buy on separate markets working conditions and safety at work, but must purchase them together with a wage from a firm. In such a setting, neither the firm nor the worker will be indifferent to the composition of the bundles traded, as neither has access to markets which would allow them to consume combinations of services different from those they exchanged on the labor market. Thus, both firms and workers have an incentive to negotiate a (possibly unique) efficient combination of, say, wages, working conditions, and work effort, whereas both parties would regard, for instance, any combination of five and ten dollar bills that added up to the same wage as an efficient combination. It is these missing markets that make implicit contracts viable in the absence of transactions costs and third party enforcement. Thus we concur with George Akerlof and Hajime Miyazaki "...the essential rationale for the existence of 'implicit' contracts need not be...risk-shifting between firms and workers. The reason for contractual arrangements could be technological,... organizational (internal labor markets) or as a response to informational impactedness" (1980, p. 328).

A second issue that has been raised with respect to implicit contract theory is the stress it places on differing degrees of risk aversion between the workers and the firm. The importance of this for the existing theory is worrying for two reasons. First, it is not clear that firms are systematically less risk averse than workers. Certainly it is hard to see why this should be the case for unincorporated firms.[4] The second problem is technical rather than economic. With missing markets for current goods, both the firms' and workers' indirect utility functions will contain more arguments than just profits and wages, respectively. However, unambiguous comparisons of risk aversion between multidimensional utility functions can only be made in very special cases, for example, where both parties' utility functions rank commodities in the same way (see Richard Kihlstrom and Leonard Mirman, 1974). As these conditions are unlikely to be fulfilled in the labor market, it is not clear how to extend technically the risk-sharing logic of implicit contract theory in this direction.

---

[2] Use of front-end loading as a means of insuring against a breach of contract was first explored in the on-the-job training/human capital formation literature. See, for example, Gary Becker (1962).

[3] For examples of the use of reputation to enforce contracts and the strength of the assumptions required see Benjamin Klein (1974), Klein and Keith Leffler (1981) and Lester Telser (1980). See Holmstrom (1981) for an example in the labor market.

Fortunately, expanding the set of missing markets to include certain current services allows implicit contracts to remain valuable even if the two parties do not differ in their attitudes to risk, or, indeed, if they are both risk neutral. The reason is that an implicit contract is a type of option, one that can only be exercised by the mutual consent of both parties, to trade in the future at terms different from those on the spot market. Moreover, this option is costless to write and has positive value, irrespective of attitudes towards risk, because there is a positive probability that both parties may wish to exercise it simultaneously. Of course, the fact that the implicit contract is an option does not preclude it from also serving as a vehicle for the trade of risk, and one would expect, in general, to see such contracts fulfilling both roles.

The paper proceeds as follows. In Section I, a simple two-period symmetric formation model is used to demonstrate an implicit contract and show that there are incentives for either the employer or the worker to breach the contract whenever the second-period contract wage differs from the spot market wage. Thus in the absence of either transactions costs or third party enforcement, implicit contracts will *not* dominate the spot labor market (Proposition 1). In Section II, an example is given to show that extending the dimensions of the compensation or service packages to include nontraded services implies that there are several realizations of the spot market different from the contractual compensation package that will not provide either party with an incentive to breach the contract. Section III provides a general theorem relating to the incentives to breach implicit contracts, analyzes the economic logic behind the result, and discusses the role of risk aversion in such contracts.

## I. Standard Implicit Contracts Will Always Be Broken

In order to fix notation and explain the power of the enforcement criticism of implicit contract theory, a simple conventional example of an implicit contract is given below.

Consider a firm which is seeking to employ one worker in each of two consecutive future periods. Let the wages it will pay in these periods be denoted by $w_0$ and $w_1$, respectively. Its profit function is linear in wages and separable across periods, and is given by

$$(1) \quad \pi = a - w_0 + \beta(a - w_1),\; a \geqslant w_0, w_1 \geqslant 0,$$

where $0 \leqslant \beta \leqslant 1$ is the firm's discount factor and $a$ is an arbitrary constant. Potential employees all have identical, separable, strictly concave indirect von Neumann-Morgenstern utility functions over the wages in the two periods. They have the choice of either working for the firm at wages $(w_0, w_1)$, or selling their labor on the spot market in the two periods. The wage in the spot market is stochastic; assume that it is independently and identically distributed across time periods with a continuous probability density function, $f(w), w \in [0, \infty)$. This function is known to both workers and the employer. The workers' expected utility from employment through the sequence of spot markets is

---

[4] In the case of corporations the risk they assume under the implicit contract is resold via the stock market. On this see Azariadis and Stiglitz (1983), and Baily.

$$EU = (1+\beta)\int_0^\infty U(\omega)f(\omega)\,d\omega, \tag{2}$$

where $\beta$ is, again, the subjective discount factor.

Returning to the employer, we see that his problem is to choose the wages he will offer $(w_0, w_1)$ to maximize expected profits (1), subject to the constraint that the worker he hires must receive the expected utility he can obtain via the sequence of spot markets (2). If we follow the literature and assume that both workers and the firm will abide by the implicit contract even in the absence of enforcement, that is, they are not fully economically rational, and that they both know that the other party will honor the contract, then the firm's problem is

$$\max_{w_0, w_1} \Pi = a - w_0 + \beta(a - w_1)$$

$$\text{subject to}\quad U(w_0) + \beta U(w_1) \geqslant (1+\beta)\int_0^\infty U(\omega)f(\omega)\,d\omega.$$

The first-order conditions to this problem, which are sufficient conditions in this case, are simply

$$\partial\Pi/\partial w_0 = -1 - \lambda U'(w_0) = 0,$$

$$\partial\Pi/\partial w_1 = -\beta - \lambda\beta U'(w_1) = 0,$$

where $\lambda$ is the Lagrangian multiplier on the constraint. Solving these shows that at the optimum $U'(w_0) = U'(w_1)$ and so the optimal wages in the two periods are the same.[5] Denote this optimal choice by $\hat{w}$. To check that this choice of a wage contract makes the firm better off than simply buying from the sequence of spot markets, we replace $w_0$ and $w_1$ in the constraint by $\hat{w}$ to get

$$(1+\beta)U(\hat{w}) = (1+\beta)\int_0^\infty U(\omega)f(\omega)\,d\omega.$$

By strict concavity (i.e., risk aversion) of $U(\cdot)$ this implies

$$U(\hat{w}) < U(\bar{w}), \qquad \text{where } \bar{w} = \int_0^\infty \omega f(\omega)\,d\omega,$$

and so $\hat{w} < \bar{w}$. Thus the wage that the employer must pay in each period is less than what he could expect to pay on the spot market, and so his expected profits will be higher if he uses the contract. Clearly, if this is the case, then we would expect labor transactions to be carried out through the use of such labor contracts and hence cause the spot markets to thin or even close, provided there is at least one future period during which labor will be traded (Holmstrom, 1983). Although the specific result in the example above, namely that workers will fully insure against fluctuations in wages, relies on the risk neutrality of the firm, provided the degree of risk aversion on the part of labor is higher than that of the firm a similar result will hold.

The smoothing out of the time path of contractual wages compared to spot market wages relies crucially on the assumption that both parties will fulfill the terms of the contract. Were the contract explicit and third-party enforcement certain, costless, and comprehensive, then this assumption would be of little concern. However, explicit labor contracts govern only a small proportion of labor market trades. In response to this objection, one can appeal to the notion of an implicit contract and claim that, although few trades in the labor market use formal contracts, the bulk use informal or implicit contracts; that is, customary, unwritten understandings as to the obligations of the two parties to the trade.

While there seems to be little doubt about the extent to which implicit, customary contracts have developed to avoid the cost of explicit contracting, such an appeal immediately raises the question of enforcement of the agreement. In the case of an explicit contract, the possibility of an appeal to a third party, for example, the courts, the NLRB or an arbitrator, that may result in damages or compelled performance, ensures that in some cases where there is an incentive to breach the contract, for instance, because the spot wage is below the contract wage, the firm will not do so. Similarly, workers will be discouraged from "cheating." Without such third-party enforcement, any incentive to cheat will in fact result in a breach of the contract.

---

[5] This result hinges on the risk neutrality of the firm and the equality of the firm's and the worker's discount rates.

Let us turn to an exact description of when the implicit contract of this section will be breached. In the second period, should the spot market wage $w$ be greater than the contractual wage $\hat{w}$, the worker has an incentive to quit and enter the spot market. Thus in the absence of transactions costs, the employee will breach the contract whenever $w > \hat{w}$ in the second period. Similarly, if $w < \hat{w}$, the employer, in the absence of transactions costs, will breach the contract by firing the worker and hiring through the spot market. In other words, only when the second-period spot wage happens to coincide with the contractual wage will the contract be carried out. As we have assumed that the probability distribution over the spot wage is continuous, this implies that, with probability one, the contract will be breached in the second period. Naturally the certainty of breach in the second period is reflected in first-period behavior. Now both parties know that they will be trading on the spot market in the last period, and consequently that their first-period behavior will have no impact on their second period utility or profits. Thus only when the spot wage in the first period equals the contract wage will neither party breach the implicit contract. If we adopt the convention that a "contract" which requires the payment of the spot wage each period is not economically a contract at all, then this argument shows that implicit contracts will be broken with probability one in all periods and so will not be observed.[6] This result is stated more formally below.

ASSUMPTION 1: There is no third-party enforcement of implicit contracts.

ASSUMPTION 2: Neither side incurs a cost of transacting via the spot market in any period.

ASSUMPTION 3: The implicit contract specifies the payment of a wage for a quantity of labor time only.[7]

ASSUMPTION 4: The workers and the firm have identical information.

ASSUMPTION 5: The probability density function of the spot wage in each period is continuous.

PROPOSITION 1: *Given Assumptions* 1–5, *implicit contracts will be broken in each of the contractual periods with probability one and so will not be observed.*

PROOF: (See the Appendix.)

One response to the problems raised by the lack of enforcement of implicit contracts has been to appeal to the value of a reputation for fulfilling implicit contracts. Presumably a reputation for breaking contracts will adversely affect the terms on which that party can trade in the future, and so provide some incentive for honorable behavior. However, the work done on this method of enforcement, both in the labor market (Holmstrom, 1981) and elsewhere (Klein; Klein and Leffler; Telser) suggests that very stringent conditions are required for this to be effective. In this connection, I might note two significant problems. First, given that a breach has occurred, both sides have an incentive to claim that the other side broke the contract, irrespective of whether they did or not. But as there is no observable contract, it is not clear how the market can evaluate these two claims. A second problem that arises in this connection is the ability to distinguish deliberate (marginal) breaches of the implicit contract from honest confusions and disputes as to the actual content of the implicit contract. All contracts are, of course, incomplete and so the interpretation of the terms and their extension to unusual events inevitably rests to some extent on unwritten understandings. In the case of an implicit contract, this goes even further with all of the terms left unwritten. In such a situation, the possibilities for honest misunderstandings seem to be very large, and such misunderstandings may well be difficult to distinguish from deliberate nonperformance of the implicit contract. However, to the extent that mistakes can be made in distinguishing misunderstandings from breach, and in the case of implicit contracts they will be easy to make, the probability that an honest firm's brand name will be incorrectly tarnished rises, and so the value of the brand name to the firm falls. Thus it seems unlikely that loss of reputation alone would act as a significant deterrent to breach of implicit contracts in the labor market.[8]

---

[6] At the general equilibrium level, this amounts to saying that Walrasian allocations involving implicit contracts are not in the sequential core. See Douglas Gale (1978) on this.

[7] The relevance of this assumption will become clear in the next section.

Given the lack of enforcement and the weakness of reputation effects, it would appear that, unless we assume that both sides of the labor contract face significant costs of trading on the spot market, we must give up the idea that implicit contracts affect individual behavior. The example contained in the next section shows this conclusion to be premature.

## II. Labor Time and Work Effort: An Example of a Viable Implicit Contract

In the previous section, it was shown that implicit contracts would not be used in the labor market because both workers and firms realize that they will always be broken. Here I provide an example to show that a simple extension of this view of the labor market in the direction of realism will allow implicit contracts to be viable alternatives to spot market transactions.

Typically the purchase of the time of a worker is regarded as the purchase of a single input. This is, of course, an abstraction. In fact, the purchase of the worker's time involves the purchase of a multidimensional input. For instance, at the most basic level one is at least buying labor time plus a level of work effort. Moreover, these two inputs, time and work effort, can only be purchased jointly for obvious technological reasons. Let us assume for the sake of this example that the effort expended by labor is continuously observable without error by the employer, and that there exists one spot labor market in which trade is costless. In this situation, the spot market for labor will clear at a "price" $(w, e)$, where $e \geqslant 0$ is a scalar measure of work effort, at which labor time supplied equals labor time demanded.

Given diversity among either the firms' or the workers' indirect utility functions and some degree of substitutability between wage and effort on the part of both, it is likely that the market-clearing price $(w, e)$ will not lie on the contract curve between a firm and some potential employees. In this case, both sides could gain by negotiating a trade at terms different from those on the spot market. Indeed, the realization of this underlies much of the internal labor market and labor relations literature. Consideration of such issues as the optimal level of working conditions or the optimal level of work effort, and how to determine such levels, can arise only when such services as labor effort and working conditions are technologically tied in to sales of labor time, and when the spot market prices of such bundles of services do not lie on the firm and its employees' contract curve.

A further observation from the internal labor market literature is that entry into the firm is restricted to a subset of jobs known as "ports of entry" (Doeringer and Piore, pp. 43–49; Williamson et al., pp. 273–75). It is at these ports of entry that competitive labor market pressures are felt most intensely. Over time, workers gain seniority and move from these ports of entry to other positions in the firm where market pressures are less severe, and where privately agreed terms of trade govern the exchange of labor services for a compensation package. Clearly, a full model of the internal labor market would present a difficult task and one only tangentially relevant to the question of the existence of a viable implicit contract.[9] Instead, I will approximate the operation of the internal labor market by assuming that workers new to the firm, that is, those in port of entry jobs, must be paid no less than the spot market price, while those in the second, and last, period of their career with the firm can trade with the firm at terms of trade other than those implied by the market. However, there is no explicit contract with the senior workers, and so both sides of the trade could if they wished trade costlessly with the spot market instead.

---

[8] As the major result of this paper shows that the incentives to breach an implicit contract are lower than previously thought, it suggests that even a weak reputation effect may be sufficient to enforce implicit contracts.

Now consider the question of whether there exists an implicit contract which the firm could offer that would make both parties better off. Proposition 1 gives a negative answer to this question in the conventional case. However, in the case under consideration here Assumption 3 is broken, that is, labor time alone is not being traded for just a wage. As we shall see, the necessarily tied-in nature of the sale of labor time and effort together with the restricted set of labor markets available implies that, even in the absence of enforcement or costs of transacting on the spot market, such an implicit contract does exist.

Let us assume that workers have identical twice differentiable von Neumann-Morgenstern indirect utility functions which are separable over time, strictly quasi concave and increasing in $w$ and $-e$. They also have a subjective discount factor $0<\beta<1$. Finally, assume that the Inada conditions hold so that the worker's optimization will always yield an interior solution. Assume that the spot market price $(w,e)$ is identically and independently distributed over time with a continuous joint probability density function $f(w,e)$ over the nonnegative orthant, $w \in [0,\infty)$, $e \in [0,\infty)$. Workers have a two-period working life over which a worker's expected utility if he trades on the spot market is

$$EU=(1+\beta)\int_0^\infty\int_0^\infty U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon$$

Turning to the employer, assume that his twice-differentiable von Neumann-Morgenstern indirect utility function, $V$, is separable over time, increasing, and strictly quasi concave in $-w$ and $e$. For simplicity, it will be assumed that his rate of discount is the same as the workers'. Further, assume that $V(-w,e)$ satisfies the Inada conditions.

To demonstrate that a viable implicit contract does exist requires two steps. First, we must confirm that if there is a positive probability that the contract will not be broken in the last period, then there will be a contract that dominates trading through the spot market. Having shown this, it must then be demonstrated that the more realistic view of what is traded on the labor market proposed in this section guarantees that indeed the contract will be honored with positive probability. Let us take the second result for granted for the moment, that is, for every possible contractual price $(\hat{w}_1,\hat{e}_1)$, $\hat{w}_1>0, \hat{e}_1>0$, there exists a set $S(\hat{w}_1,\hat{e}_1)$ of positive probability measure such that $U(w_1,-e_1)\leqslant U(\hat{w}_1,-\hat{e}_1)$ and $V(-w_1,e_1)\leqslant V(-\hat{w}_1,\hat{e}_1)$ for all $w_1, e_1 \in S$. Stated simply, assume that there is a positive probability that the contract will not be broken in the last period.

Consider an implicit contract which requires payment of the spot wage in period zero, that is, at the port of entry, and $(\hat{w}_1,\hat{e}_1)$ in period one. For this to be acceptable to a worker who realizes that it cannot be enforced it must be true that

$$\begin{aligned}&\int_0^\infty\int_0^\infty U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon\\&\quad+\beta U(\hat{w}_1,-\hat{e}_1)\iint_{\omega,\varepsilon\in S} f(\omega,\varepsilon)\,d\omega\,d\varepsilon\\&\quad+\beta\iint_{\omega,\varepsilon\notin S} U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon\\&\geqslant\int_0^\infty\int_0^\infty U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon\\&\quad+\beta\iint_{\omega,\varepsilon} U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon.\end{aligned}$$

---

[9] Akerlof and Miyazaki (p. 334) emphasize the importance of an internal labor market approach in environments where the contracts, in their case explicit, are not enforceable against labor.

This reduces to the difference in utilities

$$(3)\quad U(\hat{w}_1, -\hat{e}_1)\iint_{\omega,\varepsilon\in S} f(\omega,\varepsilon)\,d\omega\,d\varepsilon$$
$$\geqslant \iint_{\omega,\varepsilon\in S} U(\omega,-\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon.$$

As for all realizations of $(w, -e)$ in $S, U(w, -e) \leqslant U(\hat{w}, -\hat{e})$, and for some $(w, -e)$, this inequality holds strictly; we see that (3) will hold with strict inequality. Turning to the firm, by a similar argument, for the implicit contract to be acceptable it must fulfill

$$(4)\quad V(-\hat{w}_1, \hat{e}_1)\iint_{\omega,\varepsilon\in S} f(\omega,\varepsilon)\,d\omega\,d\varepsilon$$
$$\geqslant \iint_{\omega,\varepsilon\in S} V(-\omega,\varepsilon)f(\omega,\varepsilon)\,d\omega\,d\varepsilon.$$

Again, by the definition of $S$, we see that (4) will hold with strict inequality and so there does indeed exist an implicit contract which will be preferred to a sequence of spot market trades.

I have shown that if there exist "many" (strictly, a set of positive probability measure) realizations of $(w_1, e_1) \neq (\hat{w}_1, \hat{e}_1)$ such that neither party has an incentive to breach the contract, then there does exist an implicit contract which is strictly preferred to spot market trading. But does such a set of spot market realizations exist? Certainly in the example of the previous section it did not, as only the realization $\hat{w}_1 = w$, a probability zero event, was consistent with no breach of the contract. Here, however, the situation is different. To see precisely how, let us assume that an optimal implicit contract is in place, and that it specifies a second-period price of $(\hat{w}_1, \hat{e}_1), \hat{w}_1 > 0, \hat{e}_1 > 0$. As a condition for this contract to be optimal, the firm's and worker's indifference curves must be tangential at $(\hat{w}_1, \hat{e}_1)$. This may not seem surprising, as if they were not, the two parties could adjust $(\hat{w}_1, \hat{e}_1)$ in a Pareto-improving way as far as their trades in that period were concerned. But such a change will also affect the set of spot market realizations which will not provide an incentive to breach the contract and hence the probability that such a breach will occur. This raises the possibility that such a Pareto-improving change in $(\hat{w}_1, \hat{e}_1)$ from the point of view of trade in that last period might adversely affect the probability of breach and so the value of the sequence of trades via the implicit contract. Fortunately, as is shown below, this myopic Pareto-improving movement of the contract price would in fact lower the probability of the contract being broken in the last period.

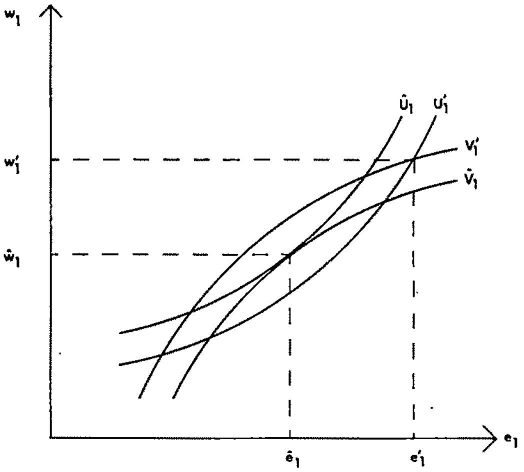


FIGURE 1

This optimal contract price $(\hat{w}_1, \hat{e}_1)$ is indicated in Figure 1, where at $(\hat{w}_1, \hat{e}_1)$, the firm's indifference curve $\hat{V}_1$ and the worker's indifference curve $\hat{U}_1$ are tangent. Notice that the assumptions made concerning the curvature of the two utility functions ensure that all such tangencies will occur in the interior of the positive orthant, and that the level sets will be strictly convex. If the spot market price realized in the last period happens to be $(w_1, e_1) = (\hat{w}_1, \hat{e}_1)$, then there is, of course, no incentive to breach the contract. But, in light of Figure 1, we see that this is not the only realization which will not lead to breach. Consider a realization such as $(w'_1, e'_1)$. Here the corresponding utility levels that would be achieved by breach of the contract are $V'_1 < \hat{V}_1$ and $U'_1 < \hat{U}_1$. Indeed, any realization of the spot market price lying on the "lens" between the two indifference curves associated with $(\hat{w}_1, \hat{e}_1)$ will not provide an incentive to breach. Notice also that if the contract had called for a last-period price, such as $(w_1', e_1')$, the set of spot market realizations that would not provide an incentive to breach that contract would be a proper subset of the no-incentive-to-breach set corresponding to $(\hat{w}_1, \hat{e}_1)$. Thus moving the contractual price from $(w_1', e_1')$ to $(\hat{w}_1, \hat{e}_1)$ would not only be Pareto improving from the point of view of last-period trade, but would also raise the probability that the contract's terms would be honored by both parties, and so all optimal contracts will specify a last-period price which lies on the contract curve.

This example shows that there will indeed be a positive probability that the implicit contract will be carried out even in the absence of transactions costs, or any third-party enforcement. In turn, this positive probability implies that there exists an implicit contract that will be strictly preferred as a vehicle of trade to a sequence of spot market trades. Moreover, risk aversion as such, as opposed to strict quasi concavity of the indirect utility functions, was not required to demonstrate this result. The economic logic behind this example and a formal proposition concerning the general case are dealt with in the next section.

## III. Missing Markets and Viable Implicit Contracts

At first sight it seems puzzling that without any barriers to arbitrage in the form of costs of trading with the spot market, and without third-party enforcement, it should nevertheless be possible to create an implicit contract which will, with some positive probability, be fulfilled. In fact, the economic reason why it is possible to construct such a contract is that certain markets for current dated goods or services are, for technological reasons, missing. This is quite apposite, as the reason for the existence of implicit contracts in the first place is the absence of a set of contingent claims futures markets. Thus, simply extending the economic logic behind implicit contracts to current dated goods markets makes such contracts viable.

More specifically, the crux of the previous example was the fact that separate markets, at any date, for labor time and labor effort are not available. For obvious reasons disembodied work effort cannot be traded except when tied in to labor time. Of course, work effort is not the only characteristic that has this feature; for example, initiative and honesty have the same quality. As a result, although labor time and work effort are substitutable in both the firm's and the worker's utility functions, they are not substitutable through the single spot market for labor time. From this arises the possibility that both parties may prefer to trade effort for labor time at a rate different than that provided by the market and so will not breach the contract. In contrast, if there were separate labor time and work effort markets, there would be a stochastic dollar value to the labor time and effort specified in the contract. Any realization of the spot prices for labor time and effort such that the market value of the quantities in the contract differed from the value of the payments specified in the contract would give one party the incentive to breach the contract in a manner exactly analogous to the example in Section I.

Although the argument above is couched in terms of the supply of tied-in services by labor, it applies in full to tied-in services provided in exchange by the firm. Certain consumption services such as job satisfaction and job safety can often be provided only by the firm. As a result of this, although the contract-specified compensation package may contain a wage/job safety pair different from that of the spot market, it may nevertheless be preferred by both the firm and the workers. Contrast this, for instance, with a wage plus refrigerator compensation package. Here, as there is a separate market for refrigerators, any realization of the spot markets for labor and refrigerators that yields a dollar value of the compensation package different to that specified in the contract will cause a breach of the contract by one party.[10]

---

[10] Notice that if, say, in the market for cars, because of tax or other reasons, the price at which the firm can buy cars differs from the price at which workers can buy

Thus, even if workers did provide just labor time, a tie in on the side of the compensation package could still result in a viable implicit contract.

Notice that the addition of these missing markets also removes the stress placed upon risk aversion and, in particular, differences in risk aversion between the parties. Here the contract provides an option to trade wages for work effort at a certain rate different from that provided by the spot market realization. However, unlike a regular option here *both* parties must jointly agree to exercise it. Nevertheless, as there is a positive probability that both will agree to exercise it, the option will be valuable to both parties and so even in the absence of risk aversion, both parties will prefer to receive costlessly the option, that is, agree on a contract, rather than not receive it and simply trade on the sequence of spot markets. However, this does not imply that the contract cannot be used to trade risk should the parties to it differ in their attitudes towards risk. For instance, in the example of the last section the degree to which the inequality (3) held measured the value of the contract to the workers. While a positive value is ensured by the convexity of their indifference curves, increasing the concavity of their utility function, loosely, raising their risk aversion, will raise the degree to which the strict inequality holds. Thus, risk can be traded through these contracts and the "degree of risk aversion" on the part of the workers will affect the value of the option and so the nature of the optimal contract.

Although the results in this paper do not rely on there being costs of transacting on the available spot markets, they do rely on the transaction cost structure specified. As in the existing literature, I have assumed that transactions costs are sufficiently high to prevent the use of contingent claims markets and have then gone further and assumed that transactions costs make trade of certain services, for example, labor effort and work safety, independent of labor time impossible. I also assumed in the example of the previous section that there was only one spot labor market. In fact, this assumption is much stronger than is required. Indeed, all that is required is that there not be a continuum of such markets. Consider the example of the previous section. If a continuum of labor markets indexed by labor effort existed, then the workers and firms would face a continuous (nonlinear) wage function increasing in work effort, and the labor market would be exactly analogous to the product market dealt with by Sherwin Rosen (1974) in his analysis of hedonic pricing. In this situation, the implicit contract will be breached in the second period unless the realization of the market wage function both ran through the point in $(e, w)$ space represented by the contract and was tangent to both parties' indifference curves at that point. Thus the existence of a continuum of markets, as Rosen's work suggests, allows labor time and effort to be implicitly traded separately and so results in the implicit contract being breached with probability one. Provided there is not a continuum of markets because of, for example, fixed costs in setting up labor markets, or nonconvexities in the firms' profit functions, then the results of Section II will continue to hold as will be shown in the proof of Proposition 2.

The final assumption made concerning transactions costs was that trades at nonmarket prices could only be carried out after the worker had been with the firm for one period. As explained previously, this assumption is made to mimic some of the findings of the internal labor market literature and, in particular, the fact that only certain jobs within firms, the so-called ports of entry, are open to outside workers, and that the labor market constrains the firm's compensation package setting behavior most severely at these ports of entry. Of course, it would be better to derive such internal labor market behavior from a rigorous maximizing model.[11] However, such a task is both non-

---

them, then there is still an incentive for intracontract trade despite the lack of a missing market.

[11] Several complementary rationales can be given for this. See Williamson et al. Perhaps the rationale most in keeping with the spirit of this paper is that new workers

trivial and tangential to the purposes of this paper, and so I have assumed that the result of such a seniority analysis can be approximated by the time path of trading specified above.

The results of the previous section and this one are summarized in Proposition 2 below, the proof of which is contained in the Appendix and is a straightforward generalization, using somewhat weaker assumptions, of the argument used in Section II.

ASSUMPTION 1: The payments by the firm to the worker are made up of two types of goods and services, those that are freely traded on competitive markets and those that are not. The former are reduced to their dollar value, summed and called the wage. The wage together with the payments of nontraded goods or services form a compensation package vector $w \in \Re^n_+$ where $n \geqslant 2$ and $w \geqslant 0$.

ASSUMPTION 2: The services provided by the worker are made up of two types, those traded on competitive markets and those that are nontraded. For simplicity and with no loss of generality, it will be assumed that only one service is tradable. The services provided by the worker are denoted by the vector $e \in \Re^m_+$ where $m \geqslant 1$ and $e \geqslant 0$.

ASSUMPTION 3: Assume both the worker and the firm have separable (over time) indirect von Neumann-Morgenstern utility functions which are strictly quasi concave, fulfill the Inada conditions, and have identical subjective time discount factors $0 < \beta < \infty$. Their utility functions are labor: $U(w, -e)$: $\Re_+^{(n+m)} \to \Re_+$; firm: $V(-w, e)$: $\Re_+^{(n+m)} \to \Re_+$. *Note*: The signs on $w$ and $e$ are reversed for the two parties to rule out freely available productive services or payments.

ASSUMPTION 4: Let there be a finite number, $s \geqslant 1$, of spot markets indexed by $i$. Denote the realizations of $(w, e)$ on each market by $\phi_i$ and the vector of market realizatons in each period by $\Phi$. Assume that (*i*) $\Phi$ is identically and independently distributed over time. (*ii*) The probability that the minimum wage realized in the vector of markets is less than or equal to any prespecified positive number is positive. (*iii*) The elements of $\Phi$ are jointly and continuously distributed according to $f(\Phi)$ such that $\Pr(w_i \ll w_j | e_i \gg e_j) = 0$ for all $i$ and $j$.

ASSUMPTION 5: Let $m, n$ or both be greater than one.

PROPOSITION 2: *Given Assumptions* 1–5, *there exists a contract which is strictly preferred by both parties to a sequence of spot market trades.*

PROOF: (See the Appendix.)

## IV. Conclusion

I have shown in this paper that a modest extension of the set of missing markets to include some characteristics of labor services or the compensation package will enable implicit contracts to remain viable even in the absence of costs of trading in the labor market or third-party enforcement. Thus, the very aspect of the economy that gives rise to implicit contracts, namely the absence of a complete set of markets, if extended a little, helps such contracts overcome the hurdle of enforcement. Moreover, this extension also enables us to do away with risk aversion as the sole explanation of implicit contracts. In view of this, this paper can be interpreted as a response to Akerlof and Miyazaki's call for further research:

> In some fundamental way, the weakness of implicit labour contract theory in explaining layoff unemployment seems to lie in its reliance on risk aversion as sole *explanans* for the existence of labour contracts. Risk aversion is obviously *one theoretical* reason for the use of contracts. However, if the im-

---

have only an incomplete knowledge of the attributes of the job and so are likely to quit after the first period (Boyan Jovanovic, 1979; Louis Wilde, 1979). Neither the new workers nor the firm will wish to invest in costly bargaining during the first period but will trade at market prices. See Armen Alchian (1969).

> plicit labour contract theory is to explain non-neoclassical aspects of labour transactions, then perhaps we should also search for a rationale for labour contracts that incorporates the (implicit!) institutional content of labour markets. [p. 335]

The major difference between this program for research and this paper is that I deal with implicit contracts which are unenforceable, whereas Akerlof and Miyazaki, terminology notwithstanding, dealt with explicit, that is, enforceable by at least one party, contracts.

While this analysis helps resolve enforcement issues in the sense of showing that the parties will often strictly prefer to trade with each other rather than the market, it does raise simultaneously a very different type of enforcement issue. While they may well prefer to trade with each other, are there any forces at work to ensure that the contracting parties will trade at the terms specified in the contract? The two parties now face each other in a bilateral monopoly situation whenever the realization of the spot market is such that they would both prefer not to trade with the market. What is to stop the firm, in the context of the example in Section II, from unilaterally altering the wage it will pay in order to appropriate more of the gains from trade? Similarly, the workers are free to set unilaterally their level of effort for the same purpose. These bilateral monopoly questions are both well known and difficult. While optimal contracts of the kind described in this paper are strong bargaining equilibria, they are not perfect equilibrium points to the bargaining game where the strategy sets are taken as the levels of the elements of the compensation package and the levels of the labor services provided (see William Vickrey, 1959, and Reinhard Selten, 1975). However, an analysis of these issues is beyond the scope of this paper. See David Kreps and Robert Wilson (1982) for an important discussion of these issues from a game theoretical viewpoint. It seems to be inevitable, though, that if the enforcement problem, in the sense of ensuring that there is a strict preference for bilateral trade between the parties, is to be confronted, one must embark on a detailed analysis not only of the resulting bargaining issues but also the effect of the prospect of future bargaining, and the uncertainty that involves on the implicit contracts struck.[12]

## APPENDIX

### *Proof of Proposition 1*

Consider a contract of arbitrary positive length $T$ and number the periods of the contract from the last period, for example, period 1 is the last period. Let the implicit contract specify the payment of a sequence of nonnegative wages $(w_0,\dots,w_T)$. Let the spot market wage in each period, $w^i$, be stochastic and independently and identically distributed. Let the common continuous probability density function with support $[0,\infty)$ be denoted by $f(w^i)$. The firm's indirect utility function, $V(\cdot)$, is separable over time and monotonically decreasing in each period's wage. The worker's indirect utility function is also separable over time but is monotonically increasing in each period's wage. The firm has a discount factor $\beta, 0 \leqslant \beta < 1$.

I proceed by induction. It was shown in Section II that the contract would be broken with probability one in the last period. Assume that it is broken with probability one in all periods 1 through $n$, $1 \leqslant n \leqslant T-1$, then the firm's expected utility at $n+1$ is

$$V(w_{n+1}) + \sum_{i=1}^{n} \beta^i \int_0^\infty V(\omega^i) f(\omega^i)\, d\omega,$$

which is monotonically decreasing in $w_{n+1}$. Thus the firm will break the contract in period $n+1$ if $w^{n+1} < w_{n+1}$. By an identical argument the worker will break the contract in period $n+1$ if $w^{n+1} > w_{n+1}$ and so, as $f(w^{n+1})$ is continuous, the contract will be broken in period $n+1$ with probability one.

---

[12] On the impact of risk and risk aversion on Nash bargaining models, see Alvin Roth and Uriel Rothblum (1982).

### *Proof of Proposition 2*

Without loss of generality, I consider the two-period case. The extension to more, but finite, periods is trivial. Note that the use of an implicit contract does not constrain either party's actions and so cannot lead to a reduction in utility either *ex ante* or *ex post* relative to a sequence of spot market trades. For a utility improving contract to exist, it is sufficient that a contract vector $(w_1, e_1)$ can be chosen such that the following set $S$ is nonnull and has positive probability measure:

$$S(w_1, e_1) = \{(w, e) | U(w, -e) > U(w_1, -e_1), V(-w, e) > V(-w_1, e_1)\}.$$

Choose $(w_1, e_1)$ on the contract curve between the workers and the firm. Consider the following two level sets:

$$\bar{U} = \{(w, e) | U(w, -e) \geqslant U(w_1, -e_1)\};$$
$$\bar{V} = \{(w, e) | V(-w, e) \geqslant V(-w_1, e_1)\}.$$

By the strict concavity of the utility functions, these sets are strictly convex on the interior of $\mathfrak{R}_+^{(n+m)}$, and there exists a tangent hyperplane to each set at $(w_1, e_1)$. Pick one such tangent hyperplane. All points on the tangent hyperplane are in the complements of $\bar{U}$ and $\bar{V}$. But the intersection of these complements is simply $S$. Thus $S$ is not empty.

The following argument proves that $S$ has positive probability measure. Consider a tangent hyperplane through $(w_1, e_1)$ and some point $(w_0, e_0) \neq (w_1, e_1)$ on that hyperplane. By the definition of a tangent hyperplane, at such point $(w_0, e_0)$ there must exist an open $(n+m)$ dimensional ball which does not contain any points in $\bar{U}$ or $\bar{V}$, i.e., lies within $S$. By Assumption 4, this ball, and so $S$, has positive probability measure.

## REFERENCES

**Akerlof, George A. and Miyazaki, Hajime,** "The Implicit Contract Theory of Unemployment Meets the Wage Bill Argument," *Review of Economic Studies*, January 1980, *47*, 321–38.

**Alchian, Armen A.,** "Information Costs, Pricing and Resource Unemployment," *Economic Inquiry*, June 1969, *7*, 109–28.

**Azariadis, Costas,** "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

______ **and Stiglitz, Joseph,** "Implicit Contracts and Fixed Price Equilibria," *Quarterly Journal of Economics*, forthcoming, 1983.

**Baily, Martin Neil,** "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

**Becker, Gary S.,** "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, October 1962, *70*, 9–49.

**Doeringer, Peter B. and Piore, Michael J.,** *Internal Labor Markets and Manpower Analysis*, Lexington: D.C. Heath and Co., 1971.

**Gale, Douglas,** "The Core of a Monetary Economy without Trust," *Journal of Economic Theory*, December 1978, *19*, 456–91.

**Green, Jerry and Kahn, Charles,** "Wage-Employment Contracts," *Quarterly Journal of Economics*, forthcoming, 1983.

**Grossman, Sanford J. and Hart, Oliver D.,** "Implicit Contracts, Moral Hazard and Unemployment," *American Economic Review Proceedings*, May 1981, *71*, 301–07.

**Hart, Oliver D.,** "Optimal Labour Contracts under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–36.

**Holmstrom, Bengt,** "Contractual Models of the Labor Market," *American Economic Review Proceedings*, May 1981, *71*, 308–13.

______, "Equilibrium Long-Term Labor Contracts," *Quarterly Journal of Economics*, forthcoming, 1983.

**Jovanovic, Boyan,** "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, *87*, 972–90.

**Kihlstrom, Richard E. and Mirman, Leonard J.,** "Risk Aversion with Many Commodities," *Journal of Economic Theory*, June 1974, *8*, 361–88.

**Klein, Benjamin,** "The Competitive Supply of Money," *Journal of Money, Credit and*

*Banking*, November 1974, *6*, 423–51.

______ and **Leffler, Keith B.**, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, *89*, 615–41.

**Kreps, David M. and Wilson, Robert**, "Sequential Equilibria," *Econometrica*, July 1982, *50*, 863–94.

**Rosen, Sherwin**, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, January 1974, *82*, 34–55.

**Roth, Alvin E. and Rothblum, Uriel G.**, "Risk Aversion and Nash's Solution for Bargaining Games with Risky Outcomes," *Econometrica*, May 1982, *50*, 639–47.

**Selten, Reinhard**, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, Spring 1975, *4*, 25–55.

**Telser, Lester, G.**, "A Theory of Self-Enforcing Agreements," *Journal of Business*, January 1980, *53*, 27–44.

**Vickrey, William**, "Self-Policing Properties of Certain Imputation Sets," in A. W. Tucker and R. D. Luce, eds., *Contributions to the Theory of Games*, Vol. V, Princeton: Princeton University Press, 1959, 213–46.

**Wilde, Louis L.**, "An Information-Theoretic Approach to Job Quits," in Steven A. Lippman and John J. McCall, eds., *Studies in the Economics of Search*, Amsterdam: North-Holland, 1979, 35–52.

**Williamson, Oliver E., Wachter, Michael L. and Harris, Jeffrey E.**, "Understanding the Employment Relation: An Analysis of Idiosyncratic Exchange," *Bell Journal of Economics*, Spring 1975, *6*, 250–78.

# Income Distribution, Lifetime Savings, and Bequests

*By* PAUL L. MENCHIK AND MARTIN DAVID*

The major purpose of this paper is to answer two questions: does aggregate consumption vary with the degree of income inequality in an economy? Does the marginal propensity to consume out of *lifetime* resources vary with one's resources? We argue that a *microeconomic* factor, the technology of the production of human wealth, lifetime labor earnings, provides the key in answering these questions that have occupied the attention of macroeconomists for so long. In addition we also examine how terminal wealth (wealth held at death) varies by age at death, conditional upon one's lifetime earnings. This issue is relevant in testing the prediction of a variant of the life cycle model of saving, a model having enormous academic support.

While at one time it was thought that equalizing the income distribution would increase consumption, the models of Milton Friedman (1957) and Franco Modigliani and Richard Brumberg (1954) imply no such distributional effect. An important paper by Alan Blinder (1975) reconsiders the effect of resource inequality on consumption in a life cycle framework.

In Blinder's model, the consumer chooses his time path of consumption $c(t)$ in order to maximize lifetime utility and is subject to the constraint that the present discounted value of both consumption and terminal wealth or bequests (if any) equals lifetime resources, $W$. Formally, the budget constraint is

$$(1) \qquad \int_0^T c(t)e^{-rt}\,dt + K_T e^{-rT} = W,$$

with $r$ the rate of interest, $t$ the consumer's age, $K_T$ the bequest, $W$ the sum of the present value of earnings and bequests received (*inter vivos* gifts are treated as discounted bequests in the model), and $T$ the certain length of life. Blinder has shown that the relationship between lifetime resources and bequests is the key to our questions. If the elasticity of bequests with respect to $W$ is constant, an elasticity greater (less) than unity implies that a reduction in inequality will, other things constant, decrease (increase) aggregate private saving.[1] The characteristics of the bequest function have recently been the subject of empirical investigation. However, the few studies that do exist have relied upon proxies for bequests, or proxies for lifetime earnings, or both. We have assembled a data base that provides *actual* bequests and *actual* income and earnings. Our data do not rely on prospective or retrospective questions about income or questions about the respondent's net worth at a point in time. Instead, we use income tax and probate record data. Although tax evasion may bias these sources, the errors are small compared with known biases, due to response and nonresponse error, in the reporting of assets in conventional surveys.

Understanding the characteristics of the bequest function has importance to a plethora of issues not discussed here. Some of these are the long-term source of inequality in income and wealth (Frederick Pryor, 1973;

---

*Department of Economics, Michigan State University, East Lansing, MI 48824, and Department of Economics, University of Wisconsin-Madison, Madison, WI 53706. We have benefited from discussions with Alan Blinder, Peter Diamond, Daniel Hamermesh, and Franco Modigliani, and have received helpful comments from participants in workshops at the universities of Chicago, Western Ontario, Queens, Michigan State, Michigan-Ann Arbor, University of California-Los Angeles, Wisconsin-Madison, Pennsylvania, and the National Bureau of Economic Research. This research was supported by National Science Foundation SOC77-27358 and by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison by the Department of Health, Education and Welfare pursuant to the provisions of the Economic Opportunity Act of 1964.

[1] If savings that were eventually bequeathed constituted a trivial share of total saving this approach would have limited importance. However, recent work by Michael Darby (1979) and by Laurence Kotlikoff and Lawrence Summers (1981) find that saving for bequest constitutes an important share of private saving.

Gary Becker and Nigel Tomes, 1979; A. B. Atkinson and A. J. Harrison, 1978; Blinder, 1976a), intergenerational economic mobility (Blinder, 1976a; Menchik, 1979; A. Shorrocks, 1979), the efficiency and incidence of a lifetime consumption tax (ourselves, 1982), whether or not government bonds are net wealth (Robert Barro, 1974), and the interrelationship between public pensions and private saving (Barro, 1978; ourselves, 1981).

## I. Motivation for Bequests

Alfred Marshall recognized a bequest motive arguing that "family affection is the main motive for saving," and that a man is interested in "leaving his family to start from a higher round of the social ladder than on which he began" (1949, pp. 227–28). As evidence supporting this bequest motive, Marshall cites the "fact" that men fail to dissave during retirement.[2] Until recently, however, the bequest motive for saving has been rather ignored.

### A. *Certain Lifespan*

Suppose individuals reproduce themselves and derive utility both from their own consumption and, in the spirit of Marshall, the economic status—or income—of their children. In such a model, parents will bequeath because they derive utility from augmenting the resources available to their children. The utility function can be written

$$U_t = U_t(C_t, W_{t+1}), \tag{2}$$

with $C_t$ the lifetime consumption of parents and $W_{t+1}$ the lifetime resources of their children.[3] We define $W_{t+1}$ as the sum of three components:

$$W_{t+1} = E_{t+1} + (1+\pi(H_t))H_t + (1+r^*)F_t, \tag{3}$$

where $E_{t+1}$ is the children's earnings capacity in the absence of any parental investments, an exogenous component determined by luck and genetic endowment. The variable $H_t$ is the amount expended by the parents in order to augment the earnings capacity of their children. This transfer has been referred to as human bequests (Blinder, 1976b) and would include parental expenditures on food, clothing, medical care, education, etc. The *average* rate of return yielded from these investments is $\pi(H_t)$, a decreasing function of $H_t$; $F_t$ is the financial transfer to children; and $r^*$ is the market rate of return on financial capital.

In this "Marshallian" model, parents derive utility from improving the "quality" of their children (borrowing the term from Becker and Tomes, 1976; 1979), interpreting quality to mean lifetime income. Furthermore, it is assumed that child quality is a normal good, which in this context implies that total transfers to children are a monotonically increasing function of parental resources,

$$H_t + F_t = t(W_t), \qquad dt(W_t)/dW > 0. \tag{4}$$

Parents will attempt to invest efficiently, choosing expenditures that yield the highest rates of return for their children. The division of parental investments into the human and financial categories, our assumption that parents are indifferent between a dollar of labor earnings or property income accruing to their children, and the monotonicity assumption, make it possible to predict the shape of the financial bequest function. Human bequests (food, clothing, health care, schooling, etc.) initially yield a higher return than financial investments. However, as the amount expended on a child increases the

---

[2] "That men labor and save chiefly for the sake of their families and not for themselves, is shown by the fact that they seldom spend, after they have retired from work, more than the income that comes in from their savings, preferring to leave their stored up wealth intact for their families" (Marshall, p. 228).

[3] This type of utility function has been used most recently by Becker and Tomes (1976, 1979) to analyze several issues, one being the quantity-quality of children decisions. It is argued that parents expend resources to improve the quality (the lifetime income) of their children and derive utility from doing so *regardless* of what the children decide to do with their enhanced income.

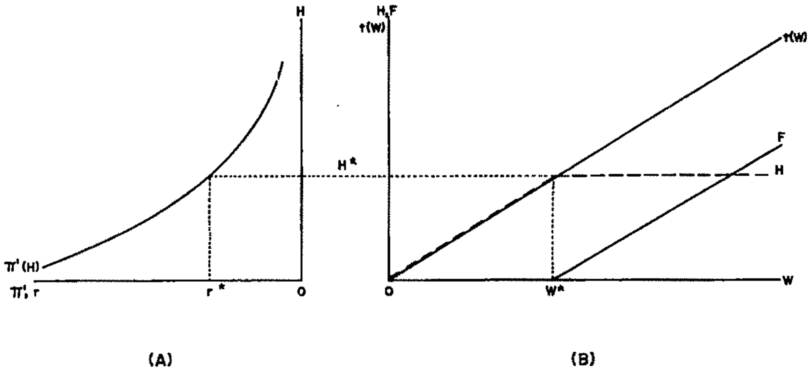


FIGURE 1

marginal rate of return, $\pi'$, falls. When the marginal rate of return on human bequests equals the market return on financial assets, all subsequent investments will be in the form of financial transfers.

In Figure 1, $H$ and $F$ are human and financial transfers, $\pi'$ indicates the varying marginal rate of return on human bequests, and $r^*$ is the market return on financial capital. Panel (A) relates the marginal return on human bequests to the amount invested. Parents will invest up to but not greater than $H^*$ in human bequests, since additional investments would yield less than $r^*$, the return yielded by financial transfers. All intergenerational transfers in excess of $H^*$ will be in the financial form. Consequently, the transfer function will appear as presented in panel (B). Human bequests will rise with parental resources, $W$, up to $H^*$ and will then become flat. Beyond $W^*$, financial transfers become positive, giving us

$$F_t = \max[0, t(W_t) - H^*]. \tag{5}$$

### B. *Uncertain Lifespan*

The most obvious defect of the model above is that in reality, one's length of life is a random variable. Individuals are uncertain about when they will die and how much of their lifetime resources need to be allocated to consumption. For a death occurring at age $s$, we partition actual bequests $A_s$ into optimal bequests $P_s$, plus accidental or unplanned bequests $U_s$,

$$A_s = P_s + U_s. \tag{6}$$

There are several reasons to believe that the expected value of $U_s$ will be positive. If individuals exhibit risk aversion toward depleting their wealth at too early an age (James Davies, 1981) we would expect individuals whose optimal bequest is zero to leave an estate unless annuity markets are perfect. However, there is reason to believe, due to asymmetric information between buyer and seller and the ensuing adverse selection, annuity markets will be imperfect. In addition, consumer durables both yield consumption services and provide a store of wealth. If the rental market for durables is imperfect, people will die holding some. Finally, borrowing constraints may truncate negative values of $U_s$, leading to a positive expected value.

In any case, whether or not the distinction between optimal and accidental bequests is important depends upon the question that is being asked. For example, even if all bequests were accidental, econometric estimates of a bequest-resources function would yield robust estimates of the bequest function as long as the world did not change in either of the following ways. First, if members of future cohorts are better at predicting their longevity than past cohorts, they could economize on accidental bequests and consume more of their resources themselves. Second, if financial institutions become better at predicting longevity as determined by personal characteristics, market failure due to adverse selection will be attenuated and accidental bequests will be reduced as people increase their purchase of annuities.

## II. Previous Empirical Studies

The bequest-resources relationship has been difficult to estimate because of serious data limitations. Attempts have been made by a number of economists (James Adams, 1980; Tomes, 1981, 1982; Kotlikoff, 1979; and Menchik 1980), using proxies in place of the independent or dependent variables. Kotlikoff uses the *Retirement History Survey* of the Social Security Administration which includes both self-reported labor earnings and net worth. Though his study uses the wealth of the living, not bequests, he shows how his wealth-earnings relationship might be the same as a bequest-earnings function estimated directly from bequest data. He finds that bequests are not elastic, but rather *inelastic* with respect to earnings, implying that the ratio of bequests to earnings fall as income rises. Several empirical problems raise questions about the study. Although lifetime earnings is the desired independent variable, Kotlikoff makes do with several years of observations. Wealth is not reduced by the costs associated with death. Another more serious problem results from errors in response to questions concerning net worth. If it is true (as seems to be the case, see Dorothy Projector and Gertrude Weiss, 1966) that high asset holdings are understated, the estimated earnings elasticity of bequests would be *biased toward zero* if no correction was made for response error. Furthermore, if high income and high wealth people are less likely to respond to financial questions than others, or if the assets favored by these people are understated (see Davies, 1979) further downward bias is introduced. Hence, the findings of bequest elasticities less than one should be taken with a degree of caution.

Both Adams and Tomes use data generated by Marvin Sussman, Judith Cates, and David Smith (1970). They both find that bequests appear to be luxury goods. The basic problem with these studies is that they do not have a measure of lifetime or annual income, and rely upon rather crude proxies. Both Adams and Tomes use the reported monthly income of the *heirs* to approximate the lifetime income of the decedents. They both regress the income measure of the heirs on the heirs' permanent characteristics, for example, education, religion. year of birth, etc. Then they use the fitted values to generate "permanent incomes" for the decedents. In spite of the substantial measurement error embodied in the income variable, statistically significant coefficients are generally estimated and the income elasticity of bequests is found to exceed unity.[4] We question whether the background variables that explain childrens' income are optimal instruments in explaining the income of parents. A data set that contained income and characteristics of people in the parents' generation would yield a more precise measure of parental income.

Menchik (1980) finds that bequests are luxury goods in a sample of Connecticut residents—the children of wealthy parents, most of whom are themselves wealthy. The data base includes the net amount the parents bequeathed to the children, and the estate of the children. Since the data do not include a measure of earnings, it was necessary to use information from another sample to estimate earnings. Using the sample generated by James Morgan et al. (1962) that included both inheritance received, earnings of heirs, and other demographic information, the elasticity of bequests with respect to lifetime income (including inheritance) is estimated to be about 2.5, implying that bequests are luxuries. Although this elasticity is significantly greater than unity at conventional levels of statistical significance, the results should be viewed with caution for two reasons. First, two *different*, though overlapping, data sets were used. Second, all the children had wealthy parents, and most of the children were wealthier than average. If the bequest function is not of the constant elasticity class, these results may not be representative of behavior at all levels of the income distribution.

---

[4] Furthermore, a replication of the Sussman et al. sample (Menchik, 1982) suggests that responses to questions concerning the amount one has inherited depart substantially from the amount that administrative records indicate one was bequeathed. Hence there may be a serious errors-in-variables problem associated with the inheritance variable in these studies.

Since the previous empirical work is inconclusive, we generated a data set with the expressed intent of estimating the earnings-bequest relationship.

## III. The Sample

### A. *Technique*

Although the conception underlying our data base is simple, the execution is not. The income tax returns were sampled using an alphabetic surname cluster sample of the persons filing tax returns in Wisconsin between 1946 and 1964 (David et al., 1974). In the years 1960–64, the list of taxpayer names was supplemented by spouses of taxpayers and beneficiaries (prior to 1964) of Social Security Accounts established for taxpayers. Studying estate wealth of these persons entailed 1) identifying the decedents, and then 2) locating probate records containing measures of the size of the estate. Step 1 was accomplished by searching records of the Social Security Administration for sample members who were marked deceased and then confirming the death through Vital Records kept in the state of death. Those taxpayers who do not have Social Security numbers were matched to the death certificate index of the State of Wisconsin.[5] Virtually all of the persons searched through the death certificate index were successfully matched to the death index.

Step 2 of the data collection entailed visits to county courthouses to locate probate records and the associated inheritance tax reports. Records associated with the decedent's wealth were located in 60.9 percent of the cases. (See Table 1.) Most of the residual matched cases (31.5 percent) can be presumed not to have a probate record, as they were not legally required to file.[6] In Wisconsin, estates of such persons are likely to be small, as those with gross estates of over $2,999 ($9,999) were required to file prior to 1973 (after May 1973). Search for probate data was limited to Wisconsin and neighboring states, and counties in Florida, California, and Arizona. The Social Security Administration indicated deaths in some other states. We did not search for probate records in those cases. In addition, Social Security records on deaths are not always consistent with Wisconsin vital statistics. As a consequence, probate records could not be searched for in about 5 percent of the decedent cases. These problems are reported in item 4 of Table 1.

The population studied differs from the universe of decedents. First, persons died during the period of observation without entering our taxpayer sample; second, some decedents entering our sample were not linked to probate or wealth information. The first problem arises because some persons who died during the period 1947–78 did not file a tax return in Wisconsin from 1946 to 1964. This group includes minors, retired and other persons whose income was primarily nontaxable transfers, and nontaxable spouses of taxpayers prior to 1960. The spouses may be gainfully employed as family workers in a farm or business, or they may have provided income-in-kind at home. One measure of the extent of this problem is that men constitute 64.4 percent of decedents matched to households with tax records.

The second problem with the data arises because persons migrating out of state are systematically excluded from the analyses presented below. Hence, any differential pattern of wealth-holding associated with migration cannot be detected. These decedents,

---

[5] In fact, the interstate network for reporting deaths implies that Wisconsin is informed of out-of-state deaths of Wisconsin residents. Only nonresident decedents must be searched for in other states.

[6] It may be that some records were missing owing to discrepancies between the place of death and the county of residence at the time of death. Great efforts were made to search all counties for probate information when the last known address disagreed with the information present on the death certificate.

TABLE 1—RECONCILIATION OF SAMPLE

| | | Total |
|---|---|---|
| 1. Decedents (of Wisconsin taxpayers or spouses) | | 4,814<br>(100.0) |
| 2. Identified Probated Estates | | 2,932<br>(60.9) |
| 1. Incomplete data | 64 | |
| 2. No income tax records for household | 79 | |
| 3. Matched complete probate records | 2,789[a]<br>(57.9) | |
| 3. Death Verified in State Vital Records for Social Security Decedents | | 1,515<br>(31.5) |
| 1. County location confirmed | 497 | |
| 2. County location not confirmed | 890 | |
| 3. Not Social Security decedent | 128 | |
| 4. Inconsistencies and Incomplete Search | | 367<br>(7.6) |
| 1. Social Security death not located in state vital records | 262 | |
| 2. Tax sample data insufficient for search or decedent not located in Vital Records | 26 | |
| 3. Social Security decedent located in a state not searched | 79 | |
| 5. Out-of-State Decedents (included above) | | 264 |
| 2.3 Matched complete probate records | 56 | |
| 3.2 Death verified in other states | 129 | |
| 4.3 Social Security decedent located in a state not searched | 79 | |

*Note:* Percentages are shown in parentheses.

[a] An additional 17 probate records could not be traced back to the original taxpayer households.

however, constitute a small proportion (5.5 percent) of the sample.

The analysis proceeds on a sample of men who filed three or more tax returns in Wisconsin and died as residents. The search process was successful in identifying wealth transfers at death for most of these persons, and the analysis will incorporate corrections for the data missing for the remainder.

## IV. Definitions of Variables and Estimations

Our dependent variable $B^*$ is the sum of three components: 1) net estate, $B$; 2) life insurance proceeds, $I$; and 3) *inter vivos* transfers (or gifts) $G$, that appeared in the probate records, discounted forward to the subject's death at a real interest rate of 1 percent. All values are stated in 1967 dollars.

The critical independent variable is a measure of lifetime earnings and inheritance received. Our income data are gleaned from Wisconsin income tax returns filed over the period 1946–64. At most, nineteen years of income occur, and an average of about thirteen to fourteen years are sampled, depending on the birth cohort (see Table 2). Two problems arise from the confounding of life cycle and cohort effects. First, since labor earnings follows a life cycle pattern, 30- and 50-year old men who earn the same amount in a given year are likely to have different lifetime earnings. The 30-year old is likely to be richer over the course of his life, and will bequeath more if bequests are normal goods. Second, if relative consumption standards change across cohorts, two people who have the same lifetime earnings but are members of different birth cohorts may differ in their willingness to supply bequests to their children.[7] One possible solution to this prob-

---

[7] Pryor hypothesizes that bequests will be a linear function of an individual's income in excess of some

TABLE 2—PERCENTILES AND OTHER STATISTICS FOR $E_i$ AND TOTAL INCOME[a]

| | Birth Year Cohort | | | |
|---|---|---|---|---|
| | 1910–24 | 1900–09 | 1890–99 | 1880–89 |
| Wage, Salary, Self-Employment Income, $E_i$ ($r=.01$) | | | | |
| Median | 5,740 | 5,056 | 3,540 | 2,416 |
| Mean | 5,705 | 5,311 | 4,074 | 2,700 |
| 80th Percentile | 7,520 | 7,115 | 5,346 | 4,013 |
| Ratio: $\frac{\text{80th Percentile}}{\text{Median}}$ | 1.31 | 1.41 | 1.51 | 1.66 |
| Mean Number of Tax Returns 1946–64 | 13.6 | 14.6 | 13.9 | 11.1 |
| Total Income ($r=.01$) | | | | |
| Median | 5,875 | 5,200 | 3,704 | 2,851 |
| Mean | 5,897 | 5,573 | 4,482 | 3,163 |
| 80th Percentile | 7,572 | 7,356 | 5,723 | 4,259 |
| Ratio: $\frac{\text{80th Percentile}}{\text{Median}}$ | 1.29 | 1.41 | 1.55 | 1.49 |

[a]Wisconsin male decedents; amounts shown in 1967 dollars.

lem would be to forecast (and backcast) the lifetime earnings of each of each man based upon his cohort and his earnings over the years of his life that we observe. We initially chose to break the sample into ten-year birth cohorts and used average annual discounted real earnings as our earnings measure.

Earnings must be discounted to some arbitrary point in the life cycle; we chose age 65, and used a real (after-tax) discount rate of 1 percent. The earnings measure is net of federal and state personal income taxes.[8] In addition, the detail of the Wisconsin tax return allowed us to divide income into labor earnings—the sum of wages, salary income, and self-employment earnings, and property income—the sum of interest, dividends, rent, and capital gains.[9] The earnings measure

---

socially determined "subsistence" income level. This subsistence level of income is not a biological subsistence level, but rather a level below which people feel they must devote all their income to own consumption instead of bequests. Modigliani (1975) allows for a bequest motive based upon one's lifetime resources relative to the cohort mean.

[8] The variable $E_i$ was adjusted to reflect *disposable* earned income. Actual taxes paid to the state of Wisconsin were reported on the return and subtracted from gross earned income; federal taxes were estimated as a ratio to the amount of Wisconsin taxes paid, and that estimate was also subtracted from earned income. As a result, earned income for persons living primarily on *rentier* incomes may be negative. The procedure used provides a lower-bound estimate for tax liability for the individual.

The ratio of mean U.S. tax liability to mean Wisconsin tax liability was computed on the 1974 sample of Wisconsin tax returns collected by the State of Wisconsin. Married couples in which the second earner accounted for more than 5 percent of Wisconsin Adjusted Gross Income were excluded, so that the ratio reflects the multiplier for persons who are able to take advantage of income splitting on the federal return to the maximum extent. The ratio was estimated for each of ten *individual* Wisconsin Adjusted Gross Income ranges as follows:

| Individual *WAGI* (1,000's) | Rate of *WI* to U.S. Tax Liability |
|---|---|
| 0–2 | .295 |
| 2–5 | .190 |
| 5–10 | .240 |
| 10–15 | .323 |
| 15–20 | .400 |
| 20–30 | .432 |
| 30–50 | .393 |
| 50–100 | .320 |
| 100–300 | .271 |
| 300+ | .290 |

[9] Clearly, the latter includes a joint return on invested capital and labor. We were unable to distinguish those elements. Measurements of self-employment income is also obscured because of tax provisions for depreciation and amortization which have tended historically to produce an underestimate of the return to enterprise. Exact derivations of both earned income and property income are given in Appendix E of David et al. (1974).

then, defined by the *CPI* (base 1967 = 100), is represented by

$$(7)\quad E_i = \frac{1}{N_i}\sum_{t=F_i}^{L_i}\frac{\left[E_i(t)(1+r)^{(BYR_i+65-t)}\right]}{CPI(t)},$$

where $F_i$ is the first year in which tax returns were filed; $L_i$ is the last; $N_i$ is the total number of tax returns for the *i*th individual; $E_i(t)$ is the amount of earned income reported for the *i*th year; and $BYR_i$ is the birth year of the *i*th person.

Table 2 displays information about the earned and total income measure for four of the six cohorts. Persons born before 1880 and after 1924 were excluded since earnings in years at either extreme of the age distribution would not be representative of lifetime earnings possibilities. Data on earnings of those born before 1880 are confounded by the retirement and endogenous saving decisions while schooling decisions confound earnings data of those born since 1924.

Unfortunately, inheritance received is not available in our data set. Although inheritance received constitutes a modest share of lifetime resources (Blinder, 1973), its likely positive correlation with lifetime earnings suggest the possibility of bias in our estimates.[10]

We include age at death, $A^*$, as a conditioning variable in the bequest equation since the level of accidental bequests should be related to longevity. It also seems reasonable to test the level of bequest against variables that may indicate a taste for leaving wealth to others. The principal contenders are measures of marital status and number of children. It may be hypothesized that there will be smaller levels of bequests for persons who never married; conversely, somewhat larger bequests may be left by men whose wives are living at the time of their death, on grounds that interspousal bequest motives as well as intergenerational motives are operative. To test for these demand factors, dummy variables were defined for persons who were never indicated as married on either tax returns or death certificates ($D_N$), and for persons whose spouses were living at the time of death ($D_M$).

The Marshallian model in Section II indicates that response to higher levels of lifetime resources will be nonlinear; optimal financial transfers increase with resources after some threshold value is exceeded. We implement a test for this type of response in the budget data by fitting a spline to the data, permitting the coefficient on earnings to change at the median level of earnings and at the 80th percentile. This specification implies that bequest-leaving behavior is a function of the *relative* level of resources of persons in the same birth cohort, as the value of the median and 80th percentile are larger for persons born in more recent years (see Table 2). Thus we define:

$$(8)\quad \begin{aligned} E_1 &= E_i && \text{if } E_i < E_{50} \\ &= E_{50} && E_i \geqslant E_{50} \\ E_2 &= 0 && E_i \leqslant E_{50} \\ &= E_i - E_{50} && E_0 < E_i \leqslant E_{80} \\ &= E_{80} - E_{50} && E_i > E_{80} \\ E_3 &= 0 && E_i \leqslant E_{80} \\ &= E_i - E_{80} && E_i > E_{80}, \end{aligned}$$

where $E_{50}$ is the median and $E_{80}$ is the 80th percentile of earned income for the cohort.

The remaining adjustment to the earnings measure permits differential bequeathing behavior for the self-employed. Reports of income for tax purposes by the self-employed are likely to understate their true lifetime resources. So long as self-employment activity is continued, assets used in the trade or business are not easily liquidated. Both reasons suggest that the level of bequests of individuals in self-employment may be higher than for persons with no self-employment. A dummy variable is set equal to one, and zero otherwise, for any individual reporting some self-employment income ($D_s$). The importance of that income in the total of earned income is measured by

---

[10] The magnitude of this bias could in principle be estimated using a data set containing both lifetime earnings and inheritance received. There is no such data set available, although the U.S. Treasury has embarked on the creation of such data using linked multiyear income tax and estate tax information. When (and if) that study is completed, we may be able to assess the bias in our results.

$$(9) \qquad Z = S_i^2 \Big/ \left[(E_i - S_i)^2 + (S_i)^2\right].$$

This oblique variable was chosen since $S_i$ was sometimes less than zero and the transformation assures that $Z$ lies in the unit interval.

The discussion above implies the following specification:

$$\begin{aligned}(10) \quad B^* = {} & a_0 + a_1 E_1 + a_2 E_2 + a_3 E_3 \\ & + a_4 D_N + a_5 D_M + A_6 D_S + a_7 Z \\ & + a_8 A^* + a_9 \lambda + \varepsilon_1.\end{aligned}$$

($\lambda$ and $\varepsilon_1$ will be explained below.)

### A. *Sampling Censoring*

Table 1 indicates clearly that reports on probated wealth were not available for all the decedents identified. For such persons data on $B^*$ is missing. However, the independent variables are observed, and the statutes governing the probating of estates give us upper bounds to the level of the gross estate of the decedent. This implies that the method proposed by James Heckman (1976) can be used to estimate unbiased values for the coefficients.

The procedure is as follows. Let $H$ be gross estate. Therefore,

$$(11) \qquad H \equiv B + L,$$

where $L$ is the liabilities of the decedent. The Wisconsin statutes provide that estates must be probated if

$$(12) \quad H \geqslant 3000/CPI(\tau^*)$$

for persons dying before May 1973

$$H \geqslant 10{,}000/CPI(\tau^*)$$

for persons dying after April 1973,

where $\tau^*$ is the date of death. Let $g(CPI(\tau))$ represent the right-hand side of (12). Then it is clear that

$$\begin{aligned}(13) \quad & Pr\,(\textit{probate record}) \\ & \qquad = Pr(H - g(CPI(\tau^*)) > 0),\end{aligned}$$

with $Pr$ denoting probability. Assuming that the expression above can be modelled by a normal distribution, we can estimate a multiple variable probit equation using the explanatory variables to give values of the expected probability for filing a probate record for every sample member. Heckman has shown that if the stochastic error term in the probit equation $\varepsilon_2$ is bivariate normally distributed with $\varepsilon_1$, then adding the Mills ratio

$$(14) \qquad \lambda = f(s)/[1 - F(s)]$$

to the specification gives unbiased estimates of the parameters ($a_j$). The standardized value of the product of the independent variables is $-s$, and their coefficients are estimated for each observation from the probit; $f(s)$ and $F(s)$ are the standard normal density and the cumulative normal distributions, respectively.

## V. Bequest Functions for Ten-Year Cohorts

Table 3 shows the percentiles of the distribution of $B^*$, the value of the estate including gifts and insurance.[11] The percentiles are probably lower bounds since persons with no report are unlikely to have died with zero net assets. The comprehensive character of the data available is indicated by the fact that more than 70 percent of the population of

---

[11] The reporting of the latter two items is truncated as up to $10,000 of insurance may be excluded from the estate, and only gifts three years prior to death are required to be reported. However, persons who reported insurance were likely to report the gross amount on the probate form, and so the amounts reported are complete, although some are missing altogether. For gifts, both under- and overreporting relative to the requirement occur. Some persons detail gifts for long periods (presumably to comply with federal estate and gift tax reporting requirements); others have undoubtedly made gifts that are completely unreported in order to evade transfer taxes.

TABLE 3—NET ESTATE DECILES, $B^*$

| Decile | Birth Year Cohort<br>1924 or later | 1910–24 | 1900–09 | 1890–99 | 1880–89 | Prior to 1880 |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 307 | 2,125 | 1,291 | 229 | 1,734 |
| 4 | 3,605 | 3,949 | 6,535 | 5,452 | 3,792 | 5,999 |
| 5[a] | 7,321 | 8,142 | 9,270 | 8,786 | 7,389 | 9,141 |
| 6 | 11,820 | 12,570 | 14,620 | 12,380 | 10,520 | 12,960 |
| 7 | 20,560 | 19,070 | 20,390 | 17,170 | 16,750 | 21,670 |
| 8 | 32,210 | 27,740 | 28,430 | 24,970 | 23,490 | 31,570 |
| 9 | 48,200 | 40,090 | 43,480 | 42,480 | 43,210 | 81,350 |
| Mean | 16,904 | 17,057 | 20,651 | 19,324 | 18,913 | 30,845 |
| $\sigma$ | 22,629 | 31,800 | 44,704 | 38,146 | 67,442 | 80,108 |
| $N$ | 96 | 238 | 464 | 699 | 528 | 153 |

*Note:* See fn. a, Table 2. The sample is smaller than the number of decedents in Table 1 because females, persons dying out-of-state, and some deaths which were verified in later searches were excluded. The 2,178 cases tabulated compared to approximately 2,950 cases in the final sample.

[a] Median.

decedents is covered by some report of their assets and liabilities at death.[12]

### A. *Estimates of the Bequest Equation*

Table 4 presents *OLS* estimates of the model presented in (10). The most robust finding is that marginal propensities to bequeath out of the top quintile in the earnings distribution are significantly positive, and the function in this range is elastic. The model provides a reasonable fit for all but the oldest cohort, for whom income earned reflects experiences after age 56 (the age of the youngest member of the cohort in 1945), and for whom the measure $E_i$ may be quite unrepresentative of earlier earnings.

Demographic influences on bequests do not appear significant. Neither the report of a surviving spouse, nor the absence of indications of marriage affect the level of bequest significantly.

Bequests appear to increase with age at death. Similar findings have been reported by others (Darby; Thad Mirer, 1980; Sheldon Danziger et al., 1982) and are damaging to the "no bequest" variant of the life cycle model of saving.[13]

The variable $\lambda$ does not have a significant coefficient, suggesting that the availability of probate data either is not significantly correlated to the error term in the bequest model (the estimated probit equations are not presented) or collinearity between $\lambda$ and the explanatory variables reduced its statistical significance.

We now turn to a detailed discussion of the coefficients for earned income and self-employment.

### B. *The Spline on Earned Income*

Inspection of the coefficients of $E_1$, $E_2$, and $E_3$ indicates a monotone increase in the coefficients for the two oldest cohorts. The coefficient of $E_1$ is indistinguishable from zero, while the coefficient of $E_3$ is significantly positive. So it appears reasonable to ask whether persons with earnings in the

---

[12] Tables 2–4 are based on a preliminary sample; regressions in Tables 5–8 are based on an updated sample including most of the cases for whom probate data were difficult to locate and verify. Table 1 reflects the final sample and is the most comprehensive. These sample differences account for the difference in coverage rates between Table 1 and Table 3.

[13] The wealth measure does not include annuities like Social Security benefits or private pension wealth. (Among members of the older cohorts private pension wealth was rare. See President's Commission on Pension Policy, 1980.) If nonannuity wealth grows with age, total wealth could never approach zero.

TABLE 4—REGRESSION MODEL OF THE LEVEL OF BEQUESTS[a]

| | Birth Year Cohort | | | |
|---|---|---|---|---|
| Variable[b] | 1910–24 | 1900–09 | 1890–99 | 1880–89 |
| $E_1$ | 2.185 | 2.826 | 1.992 | −2.396 |
| | (2.12) | (1.22) | (1.03) | (−.37) |
| $E_2$ | −.5567 | 1.008 | 2.534 | 11.26 |
| | (−.144) | (.28) | (.97) | (1.34) |
| $E_3$ | 10.67 | 9.686 | 9.489 | 19.28 |
| | (7.45) | (8.40) | (15.7) | (4.86) |
| $D_N$ | 6,012 | −11,990 | −21,510 | −18,830 |
| | (.25) | (−.65) | (−.61) | (−.73) |
| $D_M$ | 6,052 | −2,489 | 1,096 | 25,190 |
| | (.25) | (.10) | (.11) | (.88) |
| $D_S$ | 7,329 | 5,902 | −1,499 | 7,859 |
| | (1.28) | (.89) | (.37) | (.58) |
| $Z$ | 8,110 | 10,210 | 17,780 | 19,710 |
| | (.96) | (1.24) | (3.57) | (1.43) |
| $A^*$ | 545.9 | 743.2 | 516.3 | 1,722[c] |
| | (1.74) | (1.87) | (2.40) | (2.52) |
| $\lambda$ | 1,623 | 3,753 | −733.9 | 34,500 |
| | (.08) | (.16) | (.06) | (.96) |
| Constant | −30,750 | −39,640 | 30,030 | −152,100 |
| | (1.08) | (.97) | (1.43) | (2.10) |
| $\bar{R}^2$ | .316 | .213 | .408 | .058 |
| $\sigma_\varepsilon$ | 29,140 | 44,320 | 32,560 | 74,940 |
| $N$[d] | 171 | 344 | 517 | 378 |

*Note: t*-ratios are shown in parentheses.

[a] See fn. a, Table 2.

[b] See text for definition of variables.

[c] Seven persons whose date of birth was known to be prior to 1900 were included in this cohort. The value of $A^* = 0$ for these persons and a dummy variable was set. The coefficient is 152,897 and $t = 2.47$.

[d] The $N$ corresponds to the persons included in Table 3 for whom probate records were obtained.

range median to 80th percentile behave in a fashion more like the top earnings quintile or more like those below the median. That is, can the spline be reduced to a single discontinuity in the slope? Yes—only the eldest cohort indicates behavior in which the propensity to bequeath out of the earned income above the 80th percentile level is not significantly larger than the propensity to bequeath from the median to the 80th percentile level. Thus, it is fair to characterize the earnings-bequest relationship as a single kink occurring at the top quintile point, except for the eldest group where that point may occur somewhat lower in the earned income distribution. For earnings below the kink, a zero marginal propensity to bequeath can not be rejected, although the *OLS* coefficients are predominantly positive.[14]

An alternative interpretation of these results would be to take the coefficient estimates at face value and compute the marginal propensities to bequeath out of lifetime earnings. Assume workers had 40-year working lives. Since the estimated coefficients (for the three youngest cohorts) of $E_1$ and $E_2$ are

---

[14] We reestimated the model applying a 3 percent real discounted rate. This involved redefining the nodes of the spline. We obtained virtually identical values for the regressors, which was not surprising since the correlations between the values $E_1$, $E_2$, and $E_3$ under the two rates of interest are .995, .955, and .929, respectively.

TABLE 5—EXPECTED BEQUEST, ALTERNATIVE ASSUMPTIONS

| Percentile | Statistic | Birth Year Cohort | | | |
|---|---|---|---|---|---|
| | | 1910–24 | 1900–09 | 1890–99 | 1880–89 |
| A. Standard Assumptions | | | | | |
| $\leq 50$ | $E_{25}$ | 3,661 | 3,338 | 2,340 | 1,451 |
| | $B^*$ | 5,771 | 16,120 | 11,200 | −9,959 |
| | $B^*/E_{25}$ | 1.58 | 4.83 | 4.78 | −6.86 |
| 51–79 | $E_{65}$ | 6,571 | 6,081 | 4,446 | 3,178 |
| | $B^*$ | 9,841 | 22,010 | 15,880 | −3,689 |
| | $B^*/E_{65}$ | 1.50 | 3.62 | 3.57 | −1.16 |
| 80+ | $E_{90}$ | 8,992 | 8,135 | 6,663 | 4,802 |
| | $B^*$ | 24,270 | 32,930 | 30,660 | 28,290 |
| | $B^*/E_{65}$ | 2.72 | 4.05 | 4.60 | 5.89 |
| B. Death Assumed at Age 72 | | | | | |
| 80+ | $B^*$ | 35,300 | 39,430 | 31,110 | 19,480 |
| | $B^*/E_{90}$ | 3.96 | 4.85 | 4.67 | 4.06 |
| C. Death at Age 72 and Self-Employment Assumed | | | | | |
| 80+ | $B^*$ | 50,740 | 55,550 | 47,390 | 47,050 |
| | $B^*/E_{90}$ | 5.69 | 6.83 | 7.11 | 9.80 |

*Notes:* A. Individual is assumed to report earnings at the indicated percentile; $D_N = D_M = D_S = Z = 0$; $A^* = \overline{A^*}$; $\lambda = \bar{\lambda}$; B. Standard assumptions are modified so that $A^* = 72$; C. Standard assumptions are modified so that $A^* = 72$ and $D_S = Z = 1$.

around two, we can say that the marginal propensity to bequeath out of discounted lifetime earnings is .05, or 2/40, for those below the top quintile of earnings. The corresponding lifetime marginal propensity to consume would, of course, be .95. For those in the top quintile, the coefficient of earnings is around ten, hence, their lifetime marginal propensity to bequeath would be .25, or 10/40.

The model permits the bequest of self-employed persons to differ from the levels of non-self-employed by a constant, plus an amount that depends on the importance of self-employment income to the total of earned income. Thus, both the coefficients of $D_S$ and $Z$ should be considered in assessing the difference between bequests of the self-employed and the non-self-employed. In each cohort, the sum of $Z$ and $D_S$ is statistically significant using a critical region of .05. This result is consistent with two interpretations: 1) self-employed report less income relative to what they earn and pay less taxes but bequeath in the same fashion as non-self-employed; 2) for a specific endowment of earnings, self-employed have a greater capital accumulation than the non-self-employed.

Table 5 displays the implications of the regression model for the level of bequest of persons at the 25th, 65th, and 90th percentiles of the distribution of earnings. The expected bequest has been evaluated for persons who survive their spouse, who report no self-employment income, and who died at the mean age of decedents in their cohort (standard case).

For each of the cohorts, except the eldest, the expected bequest is positive for individuals at the 25th percentile of the earnings distribution. These results imply that the average propensity to bequeath is U-shaped as earned income increases. The significance of the declining portion of that curve is unclear because of nonreported transfer income. The likelihood of public (and possibly private) transfers rises as earned resources decline substantially below the median level of earned resources. Thus, the average propensity to bequeath out of lifetime resources is presumably overstated for the below-median group in Table 5.

TABLE 6—MARGINAL PROPENSITIES TO BEQUEATH, AVERAGE PROPENSITIES TO BEQUEATH, AND ELASTICITIES

| | | Birth Year Cohort | | | |
|---|---|---|---|---|---|
| Percentile | Statistic | 1910–24 | 1900–09 | 1890–99 | 1880–89 |
| A. Standard Assumptions | | | | | |
| $\leq 50$ | $\partial B^*/\partial E$ | 2.185 | 2.826 | 1.992 | −2.396 |
| | $B^*/E$ | 1.58 | 4.83 | 4.78 | −6.86 |
| | $\eta$ | .29 | .59 | .42 | .35 |
| 50–79 | $\partial B^*/\partial E$ | −.557 | 1.008 | 2.534 | 11.256 |
| | $B^*/E$ | 1.50 | 3.62 | 3.57 | −1.16 |
| | $\eta$ | .37 | .28 | .71 | −9.70 |
| 80+ | $\partial B^*/\partial E$ | 10.668 | 9.686 | 9.489 | 19.275 |
| | $B^*/E$ | 2.72 | 4.05 | 4.60 | 5.89 |
| | $\eta$ | 3.92 | 2.39 | 2.06 | 3.27 |
| B. Death Assumed at Age 72 | | | | | |
| 80+ | $\eta$ | 2.70 | 2.0 | 2.03 | 4.75 |
| C. Death at Age 72 and Self-Employment Assumed | | | | | |
| 80+ | $\eta$ | 1.88 | 1.42 | 1.33 | 1.97 |

An alternative calculation of expected bequest can be made by assuming a fixed life span for each birth cohort. This was done by calculating the expected bequests shown in the rows following the heading Age 72. The effect of the procedure is to extrapolate differences observed between persons dying earlier and later within each cohort to the year when the individual would reach age 72. The calculation does not also adjust the level of earned income by applying an expected life-time profile to the percentile position observed for the individual.

The last set of calculations in Table 5 shows expected bequest for persons whose sole source of income is self-employment, again adjusting the expectation to a uniform age at death of 72 years.

Table 6 translates the results of the model into elasticities evaluated at the 25th, 65th, and 90th percentile. The elasticities for the top quintile are all substantially in excess of unity, while those for lower quintiles are not significantly different from zero. The high values of the elasticities are somewhat attenuated under the uniform age at death assumption, and further reduced for the fully self-employed. However, even for the latter groups the elasticity exceeds unity.

### C. *Cross-Cohort Comparison*

When individuals are classified into earnings groups based on their position in the earnings distribution of that cohort, we find similar coefficient estimates for all but the oldest cohort (see Table 4). With one exception, the coefficients on the low- and medium-earnings levels are not significantly different from zero. At the same time, the coefficients of earnings in the top quntile are positive, highly significant, precisely estimated, and (with the exception of the oldest cohort) *remarkably similar to each other.*

Real productivity growth implies that absolute earnings of many in the top quintile of the 1890–99 cohort would place them in the second quintile (60–80 percent) of the 1910–24 cohort (see Table 2). If an absolute earnings hypothesis were appropriate, we would expect similar marginal propensities to bequeath for those two groups. The fact that this does *not* occur, and the similarity of coefficients found for the highest *relative* groups, is evidence that a relative earnings hypothesis is appropriate for bequests (see, for example, Pryor and Modigliani). Note that a relative earnings hypothesis for bequest saving enables us to understand why aggregate saving rates do not increase over time even through cross-section studies indicate higher saving rates for the rich than the poor.

TABLE 7–REGRESSION MODEL OF DISCOUNTED BEQUESTS OF WISCONSIN MALE DECEDENTS BORN 1890–1924

| Variable | Linear (Model 1) | Logarithmic (Model 2) |
|---|---|---|
| $E_{12}$ | 2.241 | – |
| | (3.568) | |
| $E_3$ | 9.358 | – |
| | (19.287) | |
| $LE_{12}$[a] | – | 0.224 |
| | | (3.551) |
| $LE_3$ | – | 1.815 |
| | | (8.111) |
| *DNEGE* | – | 2.206 |
| | | (3.385) |
| $D_S$ | 2,977 | 0.114 |
| | (.705) | (0.977) |
| $Z$ | 12,246 | 0.471 |
| | (3.334) | (3.251) |
| $A^*LT65$[a] | 718 | 0.0390 |
| | (2.834) | (3.915) |
| $A^*GE65$[b] | 60.0 | 0.0258 |
| | (0.222) | (2.431) |
| $BYR$[c] | 558 | 0.0258 |
| | (0.222) | (2.431) |
| $\lambda$ | −2,164 | −0.552 |
| | (0.928) | (4.996) |
| Constant | −43,754 | 4.451 |
| | (2.560) | (5.473) |
| $\bar{R}^2$ | .3212 | .1424 |
| $N$[d] | 1,064 | 1,064 |

*Note:* Amounts shown in 1967 dollars.

[a] $A^*LT65 = A^*$ if $A^* \leqslant 64$; $= 64$ if $A > 64$

[b] $A^*GE65 = 0$ if $A^* \leqslant 64$; $= A^* - 64$ if $A^* > 64$

[c] Year if birth less 1890.

[d] Retrieval of additional probate data increased the size of this sample from that used in Table 4; λ was recalculated to reflect the larger sample available.

## VI. Bequest Functions Estimated Across Cohorts

### A. *Departures from the Basic Model*

In Section V, we made the case that the problem of confusing earnings effects with intercohort and life cycle effects could be ameliorated by using four broad cohort classes. Upon reflection we realized that cohorts including ten years may not adequately model cohort effects. There could be intercohort effects within the groups shown in Table 4. At the same time, the similarity of the results encouraged us to pool the observations. We proceeded in the following way:

1) Pool the data for the 1890–1924 cohorts. (Note that the earnings of members of the oldest cohorts are endogenous with the retirement process, so members of this group were not pooled with the others.)

2) Add a continuous birth year cohort variable to our equation, $BYR = \text{year of birth} - 1890$.

3) Employ a *single* node in the spline, setting it at the estimated 80th percentile of earnings for those born in the same year.

Step 3 preserves the relativity property of the function for one-year, (not ten- or fourteen-year) cohorts. Step 2 adjusts for the fact that people who have the same earnings over the measurement period, but different *lifetime* earnings due to their different ages during the sampling period, will leave different-sized bequests.

Another modification was made to the dependent variable. Since people in the same cohort do not die at the same age, comparisons between individuals entail discounting bequests to a common point in the life cycle. That is, among persons of the same age, only bequests of equal *present* value should be considered equal size. We choose to discount bequests (using a 1 percent real rate) to the person's sixty-fifth birthday, whether he was alive then or not. Hence the dependent variable becomes discounted real bequests.

The final modification to the model affects the age at death variable. We allow the coefficient of age to vary by using a linear spline with the node at age 65. The estimated broken line allows the data to validate the inverse V-shape so popular among the proponents of the life cycle theory of saving.

The regression results (with $D_N$ and $D_M$ dropped from the regression) are presented in Table 7. We continue to get a sharp difference between the marginal propensities to bequeath on either side of the 80th percentile. It appears that on average, discounted bequeathable wealth increases with age at death until 65 and then levels out. The birth year coefficient is positive as expected and statistically significant. This implies that younger people are richer, in a life cycle sense, than older people who have identical earnings in the same year.[15]

To assure that the pooled regression does not force coefficients to a common value across cohorts, a series of Chow tests were examined for the several variants of the model, none of which include the linear birth year term. In all cases, the coefficients of the other regressors are statistically identical across all three cohorts from 1890 to 1924.[16]

Since the implications of the location of the spline node are so important, we thought it worthwhile to statistically search for the optimal spline point. To do this, we iteratively reran our model starting with the node at the 30th percentile and incrementing it by 10 up to the 90th percentile. The sum of squared errors was lowest at the 80th percentile and then rose sharply at the 90th percentile. We then refined the search, starting at the 70th percentile and incrementing the node by two points until the 90th. Serendipitously the trough occurred at the 80th percentile![17]

### B. *An Alternative Functional Form*

In order to demonstrate that we hold catholic views concerning the thorny question of functional form, we estimated a logarithmic form of the model. This form is not theoretically appropriate since people may make a zero bequest. Due to the nature of the data, there were no zero bequests to worry about. There were, however, twelve cases in which average earnings were negative, so the dummy *DNEG* was added to the regression as a shift variable, and these people were assigned one dollar of earnings. The

---

[15] Recall that the spline point is also conditioned on birth year (modification 3 in the text). This adjustment is estimated from a broken line regression of earnings on birth year, allowing for a change in slope at 1900. The logic is that all persons born before 1900 reached age 65 during the sampled income years, 1947–64. Because some proportion of these early cohorts could be expected to retire in their mid-60's, their reported earnings for the sample period reflect retirements as well as earnings capacity; for those born after 1900, reported earnings seldom reflect retirement (or disabilities). Once the broken line has been estimated, the (80th) percentile for each birth year was established from the standard error of the regression and the expected earnings conditional on that birth year. While the small samples of persons born in any one year imply that we do not observe accurate percentiles for any birth year, the percentiles estimated from the broken line were calibrated to assume a division of the entire sample of birth years at the desired (80th) percentile. Abstracting from other variables, we write the model as

$$B^* = a_0 + a_1 BYR + a_2 E + a_3 \max(E - \delta(BYR), 0)$$

where

$$\delta(BYR) = C_0 + C_1 BYR + C_2 \min(BYR - 1900, 0) + \varepsilon;$$

$$\text{and} \quad \underset{(-2.1)}{C_0 = -53{,}100}, \quad \underset{(2.2)}{C_1 = 30.2}, \quad \underset{(8.1)}{C_2 = -76.2}, \quad \sigma_\varepsilon = 3{,}330.$$

Birth year affects both the node in the earnings spline and the intercept. Testing the model in Table 4 for the 1890–99 cohort against the model that included a variable birth-year node for the income spline, we found no discernable difference in coefficients estimated. This result gave us confidence that the birth-year effect shown is not some artifact of multicollinearity in the model.

[16] A variant of the model in Table 7 was estimated to gain insight into the age-wealth relationship. The effects estimated were partitioned into age effects for those whose earnings exceed the 70th percentile and all others. This was done to test the hypothesis that one group engages in life cycle dissaving after age 65 while the other does not. (Modigliani suggested that an interaction between age and lifetime earnings may be a significant factor in our estimated age-wealth relationship.) The estimated coefficients do not support the hypothesis of differential effects for rich persons:

| | Rich (> 70th percentile) | Other (< 70th percentile) |
|---|---|---|
| *A*LT65* | \$725<br>(2.77) | \$723<br>(2.85) |
| *A*GE65* | −368<br>(0.80) | 211<br>(0.70) |

(Assuming independence for the estimated errors the $t$-ratio for the difference in savings of rich and others is 1.62. Other coefficients in the estimated model are similar to those in Table 7.)

[17] The change in $R^2$ associated with a shift in the node is small. The coefficients estimated for the models using the nodes from 70th to 90th percentiles are not significantly different from each other. The location of the node was the same in both linear and logarithmic versions.

coefficients of the age variables represent the rate of change per year of life in bequeathable wealth. As the estimated model in Table 7 reveals, bequests increase with age at death by about 4 percent per year up to age 65 and continue to rise, albeit at a reduced rate of about $2\frac{1}{2}$ percent per year after 65. These results imply that observed (or undiscounted) real bequests increase by 5 percent per year up to age 65 and by $3\frac{1}{2}$ percent per year after age 65. The elasticity of bequests with respect to earnings is not a constant. Bequests are inelastically supplied below the 80th percentile and elastically supplied above that point.

### C. *Family Size Effects*

Since our Marshallian model is based on intergenerational bequests, we must include data on number of children. Ideally we would want the number of children alive at the death of the parent. These data are not available. Instead we have obtained an impure measure of family size, the maximum number of dependents ($MAXDEP$) claimed on the income tax returns that the parents filed during the sampling period. This is an impure measure of number of children for several reasons. First, children may have left home prior to the sampling period and are therefore not counted. Second, not every dependent is a child or a spouse. Third, children who were at one time dependents of their father may predecease them. Fourth, some tax units may not include two parents during the sampled income years. As a result the number of children beneficiaries may be larger or smaller than $MAXDEP$ (less two). Impure as the variable is, we employ $MAXDEP$ in the model presented in Table 8. We use the maximum, not the average or minimum, number of dependents during the income sampling period, because we think the first problem we mention is the most serious one. In eight cases there are missing data on dependents so the dummy $DEPMISS$ is employed to retain these cases, with $MAXDEP$ set equal to zero.

Before discussing the results, we consider what effect varying family size should have on bequests in the Marshallian model. Larger families create additional opportunities for profitable human investment. Hence, increasing family size should shift the function relating the marginal rate of return on human bequests away from the origin, or to the left (panel (A), Figure 1), therefore the financial bequest function in panel (B) should shift rightward. Consequently, bequests should fall among those planning them. The presence of heterogeneity in the data may confound this result. Suppose people who have big families are different than those who have small families in the following way; the motive that leads them to have children is similar to the motive to invest in them. In such a case, it is conceivable that parents of large families will not only make larger human bequests than other parents,

TABLE 8—REGRESSION MODELS OF DISCOUNTED BEQUESTS OF WISCONSIN MALE DECEDENTS BORN 1890–1924[a]

| Variable | Linear (Model 3) | Logarithmic (Model 4) |
|---|---|---|
| $E_{12}$ | 2.404 | – |
| | (3.78) | |
| $E_3$ | 9.357 | – |
| | (19.3) | |
| $LE_{12}$ | – | 0.274 |
| | | (4.33) |
| $LE_3$ | – | 1.824 |
| | | (8.23) |
| $DNEGE$ | – | 2.493 |
| | | (3.849) |
| $MAXDEP$ | –1,057 | –0.125 |
| | (1.62) | (4.93) |
| $DEPMISS$ | –1,539 | –0.485 |
| | (0.12) | (0.973) |
| $D_S$ | 2,436 | 0.146 |
| | (0.82) | (1.27) |
| $Z$ | 12,390 | 0.492 |
| | (3.37) | (3.43) |
| $A^*LT65$ | 734.1 | 0.0407 |
| | (2.89) | (4.13) |
| $A^*GE65$ | 83.5 | 0.0284 |
| | (0.31) | (2.70) |
| $BYR$ | 660.8 | 0.052 |
| | (2.97) | (6.033) |
| $\lambda$ | –3,429 | –0.650 |
| | (1.20) | (5.84) |
| Constant | –42,990 | 4.219 |
| | (2.512) | (5.23) |
| $\bar{R}^2$ | .3216 | .1602 |
| $N$ | 1,064 | 1,064 |

[a] $t$-ratios shown are shown in parentheses.

but make larger financial bequests as well. Hence the presence of a bequest motive is compatible with either a positive or negative effect of family size on terminal wealth.

Table 8 presents a linear and a log model with *MAXDEP* included. The marginal effect of *MAXDEP* is to reduce bequests. The results imply an additional sibling will reduce the father's estate by $1,057 in the linear model, and by about $12\frac{1}{2}$ percent in the logarithmic model.

## VII. Conclusions

The results fail to show individuals deaccumulating wealth in old age. Indeed some results show the opposite. It is hard to understand how or why people could be accumulating in old age. One possible reason for *how* they accumulate is that they drastically reduce consumption in old age (Daniel Hamermesh, 1982; Danziger et al.). Two reasons *why* they do it comes to mind. First, the bequest motive is strong, and may become stronger, as they age. Second, persons become more and more averse to risk as they age. In any case, the data fail to support the positive predictions of the "humped" age-wealth profile.

The marginal propensity to bequeath out of lifetime earnings is greater for the rich (those in the top quintile) than others. Hence, redistribution of income from the rich to others (if expected and not transitory) will reduce lifetime saving as is defined here. Interestingly, redistribution within the bottom four quintiles will not significantly reduce lifetime saving. In general, we can say that an egalitarian state with most people at the median would likely save less than a less egalitarian state with many people at the tails of the earnings distribution.

Some caveats are in order. First, it is assumed by us (and by Blinder) that all people have the same utility function so life cycle saving (saving for future consumption) cannot be affected by the income distribution. If, however, the subjective rate of time preference is correlated with earnings, income distribution would also affect life cycle saving. Second, if the net of tax rate of return to financial wealth is related to earnings, this could alter the interpretation of our findings. If higher earners received higher net yields, findings like ours would be contaminated by a price effect falsely interpreted as an income effect. Finally, and most importantly, suppose the Marshallian model is correct. Redistribution from the rich to others would reduce financing saving, but, if human capital expenditures rose, there would be an offsetting effect on total capital supply. Indeed, our model requires the rate of return on human capital expenditures to be greater than, or equal to, the yield on financial capital. Hence a diversion of one dollar from bequests to human capital expenditures would likely to increase *total* capital stock, an outcome explicitly stated by Marshall (pp. 229–30).

## REFERENCES

**Adams, James D.,** "Personal Wealth Transfers," *Quarterly Journal of Economics*, August 1980, *95*, 159–79.

**Atkinson, A. B. and Harrison, A. J.,** *Distribution of Personal Wealth in Britain*, Cambridge, England: Cambridge University Press, 1978.

**Barro, Robert J.,** "Are Government Bonds Net Wealth?" *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

______, *The Impact of Social Security on Private Saving*, Washington: The American Enterprise Institute, 1978.

**Becker, Gary S. and Tomes, Nigel,** "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, August 1976, *84*, Part 2, S143–62.

______ **and** ______, "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *Journal of Political Economy*, December 1979, *87*, 1153–89.

**Blinder, Alan S.,** "A Model of Inherited Wealth," *Quarterly Journal of Economics*, November 1973, *87*, 608–26.

______, "Distribution Effects and the Aggregate Consumption Function," *Journal of Political Economy*, June 1975, *83*, 447–76.

______, (1976a) "Inequality and Mobility in the Distribution of Wealth," *Kyklos*, No. 4, 1976, *29*, 607–38.

______, (1976b) "Intergenerational Transfers and Life Cycle Consumption," *American Economic Review Proceedings*, May 1976, *66*, 87–93.

**Danziger, Sheldon et al.**, "The Life Cycle Hypothesis and the Consumption Behavior of the Elderly," Discussion Paper No. 697, Institute for Research on Poverty, 1982.

**Darby, Michael R.**, *The Effects of Social Security on Income and the Capital Stock*, Washington: The American Enterprise Institute, 1979.

**David, Martin, Gates, William, and Miller, Roger F.**, *Linkage and Retrieval of Microeconomic Data*, Lexington: D.C. Heath, 1974.

______ **and Menchik, Paul L.**, "The Effect of Social Security on Lifetime Wealth Accumulation and Bequests," Discussion Paper No. 671, Institute for Research on Poverty, 1981.

**Davies, James**, "On The Size Distribution of Wealth in Canada," *Review of Income and Wealth*, September 1979, *25*, 237–59.

______, "Uncertain Lifetime, Consumption, and Dissaving in Retirement," *Journal of Political Economy*, June 1981, *89*, 561–77.

**Friedman, Milton**, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.

**Hamermesh, Daniel**, "Consumption During Retirement: The Missing Link in the Life Cycle," Discussion Paper No. 930, National Bureau of Economic Research, July 1982.

**Heckman, James**, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, Fall 1976, *5*, 475–92.

**Kotlikoff, Laurence**, "Estimating the Elasticity of Bequests with Respect to Lifetime Resources from a Sample of Potential Decedents," unpublished manuscript, University of California-Los Angeles, 1979.

______ **and Summers, Lawrence**, "The Rôle of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

**Marshall, Alfred**, *Principles of Economics*, 8th ed., New York: Mcmillan, 1949.

**Menchik, Paul L.**, "Intergenerational Transmission of Inequality: An Empirical Study of Wealth Mobility," *Economica*, November 1979, *46*, 349–62.

______, "The Importance of Material Inheritance: The Financial Link Between Generations," in James D. Smith, ed., *Modeling the Distribution and Intergenerational Transmission of Wealth*, Chicago: University of Chicago Press, 1980, 159–85.

______, "Unequal Estate Division in Cleveland: Is it Altruism, Reverse Bequests, or Simply Noise?" unpublished manuscript, Michigan State University, 1982.

______ **and David, Martin**, "The Incidence of a Lifetime Consumption Tax," *National Tax Journal*, June 1982, *35*, 189–203.

**Mirer, Thad**, "The Dissaving Behavior of the Aged," *Southern Economic Journal*, April 1980, *46*, 1197–205.

**Modigliani, Franco**, "The Life Cycle Hypothesis of Saving Twenty Years Later," in M. Parkin, ed., *Contemporary Issues in Economics*, Manchester: Manchester University Press, 1975.

______ **and Brumberg, Richard**, "Utility Analysis and the Consumption Function: An Interpretation of Cross-section Data," in K. K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick: Rutgers University Press, 1954.

**Morgan, James et al.**, *Income and Welfare in the United States*, New York: McGraw-Hill, 1962.

**Projector, Dorothy and Weiss, Gertrude**, *Survey of Financial Characteristics of Consumers*, Washington: Federal Reserve Board, 1966.

**Pryor, Frederick**, "Simulation of the Impact of Social and Economic Institutions on the Size Distribution of Income and Wealth," *American Economic Review*, March 1973, *63*, 50–72.

**Shorrocks, A.**, "On the Structure of Intergenerational Transfers Between Families," *Economica*, November 1979, *46*, 415–22.

**Sussman, Marvin, Cates, Judith and Smith, David**, *The Family and Inheritance*, New York: Russell Sage, 1970.

**Tomes, Nigel,** "The Family, Inheritance and Intergenerational Transmission of Inequality," *Journal of Political Economy*, October 1981, *89*, 928–58.

______, "On the Intergenerational Savings Function," *Oxford Economic Papers*, March 1982, *34*, 109–34.

**President's Commission on Pension Policy,** *An Interim Report*, Washington: USGPO, 1980.

# The Economics of Invention Incentives: Patents, Prizes, and Research Contracts

*By* BRIAN D. WRIGHT*

Though public intervention in the market for research is virtually universal, economists have paid surprisingly little attention to the choice of the form of research incentive in a given market structure. Many studies concentrate on patents, but any assumption of their superiority over other incentives has been founded on intuition rather than on formal analysis.

In this paper I analyze the choice between three of the most common alternative means of public intervention in the research market, namely, patents, prizes, and direct contracting for research services. I show why, and under what conditions, any one of the three may be preferred by a social welfare-maximizing administrator in a competitive economy, using a model that, for the first time, pays explicit attention to differences in the informational roles of each of these alternatives.

In the extensive literature on the economics of patents (see Arnold Plant, 1934; Fritz Machlup, 1958; Charles Taylor and Z. A. Silberston, 1973; Morton Kamien and Nancy Schwartz, 1975; and F. M. Scherer, 1977, for valuable surveys), formal analysis weighs the benefits of patents as a solution to the market failure associated with the inappropriability of knowledge against the welfare cost due to the restriction on the use of the knowledge generated, and this tradeoff is optimized by patent life adjustment in William Nordhaus (1969). Scant analytical attention is paid to alternative incentive mechanisms. (An exception is Ben Yu, 1981, who considers the role of prior contracting for inventions.)

But as many writers (for example, Dan Usher, 1964; Yoram Barzel, 1968; Joseph Stiglitz, 1969; Carole Kitti, 1973; Glenn Loury, 1979; Partha Dasgupta and Stiglitz, 1980a) have pointed out in various contexts, the incentive offered by an unlimited patent to competitive researchers may be excessive, due to the "common pool problem" discussed further in Section I below. If the patent administrator and researchers share the same information, as implicitly assumed in previous models, then the patent life limitation can be adjusted to provide the optimal patent incentive, given the common pool problem. But in all such models, patents would not be chosen in a fully optimized fiscal system. Researchers and the administrator are assumed to have identical information about the shadow price of potential inventions; a patent is just a means of turning this shadow price into a monetary reward. But monetary compensation can instead be offered directly to researchers by the state. Assuming that patent revenues incur a higher deadweight loss than an equivalent amount of public funds financed by less distortionary means (for example, a minimally efficient tax system), appropriate prizes or government contracts are socially preferable to patents with optimal lives.

If the patent is ever to be the optimal incentive mechanism for research, it must possess advantages not captured in existing models. Informal discussions of patents emphasize their informational role. To include the latter as a justification for decentralized invention incentives, I incorporate an *ex ante* imbalance of information about costs and benefits of research in the model presented in Section II. But this alone is not quite enough. It is further necessary to specify that the terms of the award must be fixed before this informational imbalance is resolved, if it can ever be economically resolved. Several arguments for this assumption which is maintained throughout are presented in Section V; if it does not hold, any rationale presented here for choosing patents over other incentives with lower excess burden collapses.

---

*Department of Economics, Economic Growth Center, Box 1987 Yale Station, Yale University, New Haven, CT 06520. I thank, with the usual caveat, Marguerite Alejandro-Wright, Cindy Arfken, Martin Baily, Nuong Brennan, Steven Englander, Robert Evenson, Richard Levin, Richard Nelson, Susan Rose-Ackerman, Denis Wright, and two referees for assistance of various kinds.

The informational asymmetry of the model in Section II does make a patent system a relevant choice, in contrast to previous models. But it also transforms the common pool problem, previously viewed as correctable, into such a serious drawback for prizes and patents that contracts may be a superior alternative. Hence any of the three alternatives considered may be the best choice given the three allocative difficulties recognized in the model—lack of appropriability of knowledge, deadweight loss of the patent, and the common pool problem. The relationship between the parameters of the model and the choice of incentive mechanism is analyzed in Section III and illustrated for a simple case in Section IV. Some generalizations of the results are discussed in Section V, and conclusions follow in Section VI.

## I. Resource Allocation in a Simple Competitive Invention Model where All Agents have Identical Information

I consider the search for a specific process invention which, by lowering the cost of production of some competitively produced commodity, could produce a social benefit of value $B$. The invention is discrete (a "pot of intellectual gold") and its discovery imparts no knowledge that could be useful in the development of other technologies.[1] A public administrative authority, responsible for research activity, seeks to maximize a risk-neutral individualistic social welfare function through intervention in the market for research. At time 0, the administrator announces the type and terms of the incentive offered to researchers, and researchers make their allocative decisions. All research projects take a fixed time to complete, and finish at time $I$, when the results of the research are revealed, and the invention, if achieved, is utilized and yields its social benefit. Discounting of costs and benefits is implicit; all prices are expressed as time $I$ present values.

Research activity is carried out by a large number of small independent profit-maximizing, price-taking research firms with identical technologies. One of the factors which they employ, "research skills," measured in efficiency units, has a nondecreasing marginal cost to the research industry, and all the other factors used by researchers are in perfectly elastic supply.[2] For any firm $i$ the total cost of producing $\mu_i$ units of research activity is $c(\mu_i, m)$, where $m$ is the aggregate research produced by all firms. Because of the indivisibility of the services of the owner-manager, marginal cost $\partial c/\partial\mu_i \equiv c_1(\mu_i, m)$ first decreases then increases with $\mu_i$; minimum average cost occurs at output $\mu_i^*$, independent of $m$. For simplicity it is assumed $\mu_i^*$ equals unity, so with free entry each firm in the research industry performs just one project and the number of firms equals the number of projects $m$.[3]

---

[1] By considering a process invention, I ignore effects of different incentives on product variety which could be important for product inventions, as one referee has noted. Further, the assumption of a single goal greatly simplifies the analysis by ruling out two phenomena related to the disclosure inherent in the patent incentive. The first is that competitors can often "invent around" a patent, lowering its value to the inventor (Machlup, p. 50). This possibility in turn tends to broaden the scope of claims made by inventors, and to encourage defensive "sleeping patents" (Richard Gilbert and David Newbery, 1982). It can also make the alternative of attempting to maintain secrecy more attractive. The second is that inventions often provide some of the information needed to develop other, possibly quite different, socially beneficial inventions, which may not affect the private value of the first discovery. The additional social return to the invention will not be reflected in the value of the patent.

[2] Research skills may be owned by the researcher firms themselves. The supply situation is analogous to that of competitive farmers who own or rent land which has an upward-sloping supply curve to the industry as a whole.

[3] It is assumed that $\mu_i^*$ is sufficiently small to justify treatment of aggregate research $m$ as if it were a continuous variable. The equilibrium result, that all firms operate at optimal efficiency, is derived in Dasgupta and Stiglitz (1980a, pp. 17–18).

For the industry as a whole, the cost of $m$ units of research activity is $C(m) \equiv \Sigma c(\mu_i)$ and the industry research activity supply curve is $C'(m) > 0$ with $C''(m) \geqslant 0$. In the search for the invention, I assume that if duplication of research tasks occurs, it must be unforeseeable *ex ante* to both researchers and the administrator. (Thus a possible advantage of centrally controlled research contracts in reducing duplication is ignored.) The probability that the invention will be found (i.e., that at least one research project is successful) is $P(m)$, with $P'(m) > 0$ and $P''(m) < 0$.

A completely decentralized market will not in general yield the optimal level of research in this model. If secrecy is ruled out,[4] the invention, being available at no cost to all users, will have no commercial value. Therefore government intervention is needed to provide private research incentives. The socially optimal allocation of research effort depends upon the information available to the administrator and researchers. In this section it is assumed that the government and private researchers have identical information. They know the cost and probability-of-success functions, and have rational expectations about $m$ prior to the conduct of research.

Then assuming lump sum taxation is possible (or the marginal welfare cost of general revenue is negligible) and there are no market distortions in other sectors, the socially optimal level of research activity $m_0$ is the level at which expected marginal social benefits $BP'(m)$ equal marginal social cost $C'(m)$. This can be achieved in several ways. The administrator can announce that he will buy $m_0$ units of research at a price determined by competitive bids. Or, if $C''(m) > 0$, he can offer to pay the *ex ante* expected marginal product, $B \cdot P'(m_0)$ for any units of research offered. Then (assuming no alternative demand for research) in equilibrium

$$(1)\quad C'(m_0) = \mu_i^{-1} c(\mu_i, m_0) = c_1(\mu_i, m_0) = \partial\left[\mu_i \cdot B \cdot P'(m_0)\right]/\partial \mu_i = B \cdot P'(m_0),$$

and the optimal level of research output $m_0$ is achieved with $\mu_i = 1$ for any firm $i$ conducting research. Note that the competitive researcher ignores the marginal effect of his own research activity $\mu_i$ on $P'(m)$, consistent with the familiar price-taking behavior of an atomistic competitor.

Alternatively, the administrator might consider rewarding successful projects. This can be done by offering patent protection for a successful inventor, or by offering a prize.

Absent any deadweight loss, it might seem that an optimal patent policy would protect the full value of the invention, (see, for example, Nordhaus), and that a prize of equivalent present value would also be optimal. However, in a competitive model either incentive will create a welfare loss in the invention market if an award is made whenever success is achieved and, in the case of multiple successes, is shared equally by the successful projects or allocated randomly to one of them. Since each of the $m$ projects is considered equally likely to discover the invention, the researchers, having rational expectations, consider each project to be one "ticket" in the "invention lottery," the expected return on each being $m^{-1}B \cdot P'(m)$. Producer $i$ equates this expected return to his marginal research cost $c_1(\mu_i, m)$, so that

$$(2)\quad C'(m) = c_1(\mu_1, m) = \partial\left(\mu_i \cdot m^{-1} B \cdot P(m)\right)/\partial \mu_i = m^{-1} B \cdot P(m).$$

Since $P''(m)$ is negative by assumption, $m^{-1}P(m)$ is strictly greater than $P'(m)$, so a comparison of equation (2) with (1) shows that if the patent or prize offers the full social value of the invention, excessive resources are allocated to research. In equilibrium, the only net social gains expected from research are the rents accruing to the research skills employed by research firms; if these skills are in perfectly elastic supply, the entire research effort produces no net expected benefit at all. The problem is equivalent to that noted by H. Scott Gordon (1954) with respect to competitive fishing in an open access fishery;[5] it is, borrowing Dasgupta and Stiglitz's terminology, denoted the "common pool problem" below.

---

[4] Secrecy is technically feasible only for the limited class of inventions in which exploitation does not imply revelation of the secret. It should be clear from the discussion below that secrecy is for several reasons socially undesirable in the full information model of this section. See Taylor and Silberston, ch. 9; Edmund Kitch (1977); my 1981 paper; and Steven Cheung (1982) for further discussion of the public and private merits of secrecy as a means of protecting inventions.

The common pool problem can be solved by reducing the value of the patent or prize award to a fraction $Y$ of the social value of the discovery, where $Y$ equals the ratio of expected marginal to expected average benefit of research activity. In a multiperiod context, the reduction in patent value would be equivalent to the existing patent life limitation in the United States, except that the length of life would not be fixed at seventeen years, but would be adjusted *ex ante* so that the expected present value of the award is reduced by a fraction $(1-Y)$. Patent life reduction as a solution to the common pool problem is presented in Kitti.

Assuming intertemporally separable patent demands, the analytical equivalent of this patent life limitation in my one-period model is achieved by randomly allowing the award of a patent, given success, only with probability $Y$ which is a function of the parameters of the problem. (A tax on the patent has a greater expected welfare loss.) Except where otherwise indicated, in this paper "patents" denote incentives optimally adjusted in this way.

In this starkly simple static model, there is another way of overcoming the common pool problem with patents and prizes, which, though of dubious practicality, will provide us with a useful yardstick in following sections of this paper. It involves a change in the rules of awarding the patent right or prize. The misallocation illustrated above arises because the expected private benefit from a unit of research exceeds its marginal social benefit. When does a unit of research have a positive marginal social benefit *ex post*? Only when it results in success *and* all other units fail. If more than one researcher discovers the single discrete invention considered here, then the *marginal* contribution of *each* is zero. Hence, the requisite rule change is that the prize or patent right be awarded only in the case of a single, unduplicated success in achieving the research goal at time $I$.[6] Since $P(m)$ is defined as the probability that *at least one* research project succeeds, each researcher, having rational expectations as to $m$, will perceive the incentive to be $B \cdot P'(m_0)$, which is exactly the same as the optimal public payment for research described above.

In this model the principle is implemented by mandating that to qualify for an award, all researchers must submit their results in confidence to the planner at the end of the invention period. Competition should rule out any communication between successful applicants which would reveal multiple successes. For want of a better term, we call awards under this rule "marginal patents" and "marginal prizes." In a more realistic dynamic intertemporal context, this method of making awards presents serious administrative challenges. How is parallel invention to be distinguished from copying? Over what time interval can the occurrence of similar inventions be labeled "simultaneous" and no award allowed? How soon would the invention have been found by another researcher? Such difficulties would ordinarily preclude real world implementation of marginal patents and prizes unless, for some exogenous reason, competition must end and implementation begin at a fixed point in time. However, marginal patents and prizes are useful analytically because, unlike the alternative of reducing the reward, they yield conventional welfare results regarding patents even when informational asymmetry is introduced. They serve as yardsticks which help show, in Sections III and IV below, the extent to which recognition of the common pool problem reduces the value of the more realistic decentralized alternatives under various parameterizations of the model to be introduced in Section II.

---

[5] The application of Gordon's insight to research was noted by Usher. In an intertemporal context the benefits are dissipated in the "race to be first" (Barzel). Others who have recognized the problem in various guises include Kitti, Loury, and Dasgupta and Stiglitz (1980a). Recently Kitch and Yu have assessed patent incentives from an institutional and contractual viewpoint. In the form of an employment contract, prior contracting is much like my "contracts" option (for example, Yu, p. 221). If prior contracts are for patents, moral hazard problems regarding self-identification and motivation of researchers could be severe under informational asymmetries.

[6] That is, research is rewarded only if it makes a marginal difference, *ex post*. The method is in this sense analogous to the assessment of the "Clarke tax" and related mechanisms for obtaining honest revelation of preferences for public goods (Edward Clarke, 1971). I am indebted to Susan Rose-Ackerman for pointing out this analogy.

Though the above alternative eliminates the common pool problem for patents in the model used in this section, it has no relevance for identification of the best invention incentive system because its associated deadweight loss renders it inferior to direct payments by the public in the form of prizes or contracts.[7] Justification for using patents must be sought in a different model.

The rich informal literature on the patent system emphasizes the importance of the patent as a decentralized decision-making device, implying that information and its distribution are major elements in the rationale for the patent instrument. This rationale is explored in the model outlined below, in which the research administrator possesses information different from that available to private sector researchers. The model is used to evaluate the three alternatives which have widely used, real world counterparts: patents, prizes, and research contracts.[8]

## II. The Invention Model with Asymmetric Information

This model is an extension of that above, modified to make information about costs and rewards asymmetric. Total industry research cost is now subject to a random multiplicative disturbance $\theta$, with zero mean and finite variance $\sigma_\theta^2$. This cost function is denoted $K(m,\theta)$, where $K(m,\theta)=(1+\theta)C(m)$, with $K_1\equiv\partial K/\partial m>0$, $K_{11}\equiv\partial^2K/\partial m^2\geqslant 0$. The social payoff if research is successful, $B$, is subject to an independent random disturbance $\zeta$ with zero mean and finite variance $\sigma_\zeta^2$. Consistent with Section I, the probability that one or more researchers find the single objective is $P(m)$, with $P'(m)<0$, $P''(m)>0$.

As discussed above, an imbalance of information between the public research authority or administrator and the innovating agents is essential if a patent system is to be a candidate for the best incentive choice in a competitive model. Reasons for such an imbalance have been advanced by Kenneth Arrow (1962), Martin Weitzman (1974), Richard Nelson (1982), and many others. The assumption that information is costly to acquire, and the implication that it might not always be optimal (or even possible) for the administrator to find out everything relevant to his decision that the private researchers know, seem especially appropriate in the case of inventive activity.

The informational asymmetry is represented as follows. At time 0, when decisions are made on the level of incentive and on the allocation of resources, the researchers know $\zeta$ and $\theta$ which affect benefits and costs, respectively, but the administrator knows only the distribution of $\zeta$ and $\theta$. For example, $\zeta$ might represent a marketability factor that is known only to private firms familiar with marketing, while $\theta$ might reflect industry-specific research experience. Processes which could directly transfer information about these random variables between researchers and the administrator are assumed to be too costly to be feasible.

---

[7] Stiglitz presents a similar argument. Ward Bowman (1973, ch. 5) points out that price discrimination by the patent holder would reduce the excess burden, but he does not claim that perfect price discrimination is feasible. Assuming all price discrimination is ruled out, the welfare cost caused by a patent on a cost-reducing invention can be described as follows. For simplicity, I assume that the invention is of the run-of-the-mill type (Scherer's 1965, and Nordhaus' terminology) in which the absolute values of the cost reduction and the demand elasticity for the product are not so large that revenue-maximizing output exceeds the preinvention competitive supply, $Q_0$. (For a clear discussion, see Nordhaus, pp. 70–73, or Dasgupta and Stiglitz, 1980a). The invention lowers marginal (and average) cost from $MC_0$ to $MC_1$. A profit-maximizing patent holder maintains output at $Q_0$ rather than at its new competitive level $Q_1$. Patent revenue and the approximate welfare cost of the patent, are $Q_0(MC_0-MC_1)$ and $\frac{1}{2}(Q_0-Q_1)(MC_0-MC_1)$, respectively.

[8] Obviously these are not the only means which can conceivably be used to foster research. I have, for example, ruled out complicated combinations of patent lives and subsidies or taxes mentioned in the literature (Kitti, p. 31; Nordhaus, p. 89), and secrecy (see fn. 4 above). An auction system is presented by Barzel and used by Kitti, but "only as a standard of comparison for the current patent system. It is not being promoted as a feasible alternative" (p. 68).

## III. A Comparative Evaluation of Patents, Prizes, and Contracts

The evaluation criterion is the expected net differential social benefit relative to a contract by the administrator for $m_0$ units of research, the optimal contractual allocation, given that the administrator cannot observe the realizations of $\zeta$ and $\theta$. Two types of award are considered initially.

[1] *Patent.* The patent award is given to one of the successful discoverers at time $I$, (or shared by all successful discoverers, who exploit it as a monopoly), so that the competitive individual inventor responds to the average expected benefit. The welfare loss caused by an *unlimited* patent, which is incurred only when a patent is awarded, is defined as $L$, where the magnitude of $L/B$ is assumed at most to be of the same order as $\sigma_\zeta^2$. The common pool problem discussed above is alleviated by allowing the patent award, given success, only with some probability $Y$, less than unity. As discussed above, this procedure is the atemporal equivalent of limiting patent life; it maintains an equivalent relation between the private value of the patent and its deadweight loss.

In the model of Section I, the optimal allocation $m_0$ could be achieved by offering to purchase research inputs at the price,

$$(3) \qquad B\cdot P'(m_0)=Y_0BP(m_0)/m_0,$$

where $Y_0$ is equal to $h(P)$, the elasticity of total probability that at least one research project succeeds, evaluated at $m_0$,

$$(4) \quad Y_0=h(P)\equiv m_0P'(m_0)/P(m_0).$$

In the case of a patent awarded in the current model with asymmetric information, the value of $Y$, which is optimized by the administrator *ex ante*, is a function of the excess burden $L$, as well as the other parameters of the problem.

[2] *Prize.* If a prize is chosen, it is awarded whenever success is achieved. If several projects are successful, it is shared by all or a subset chosen at random. The prize value is optimally reduced *ex ante* to counter the common pool problem. I assume that the revenue needed to finance the prize can be raised with negligible excess burden.

The method of evaluation of patents and prizes proceeds as a generalization and extension of the approach used by Weitzman (1974) to study the choice of prices vs. quantities in controlling pollution.[9] First consider the simple example in Figure 1, in which the benefits are nonstochastic and equal to $B$. The marginal cost of research activity, $C'(m)$, is subject to a multiplicative shifter $(1+\tilde{\theta})$, where $\tilde{\theta}$ equals $+\hat{\theta}$ or $-\hat{\theta}$, each with a probability of one-half. Probability of one or more successes is a nonlinear function $P(m)$, and $BP'(m)$ is the marginal expected return function.

Under a prize option, the administrator optimizes the fraction $Y$ of the social benefits incorporated in the award. To simplify the graphical exposition, it is assumed here that $Y$ is fixed at $Y_0$; the full optimization (contained in the Appendix) confirms that this is an adequate approximation for small values of $\sigma_\theta^2$ and/or $\sigma_\zeta^2$. Researchers respond by producing $m_3$ or $m_4$ units of research. The expected differential net social gain from choosing the prize option, rather than contracting for $m_0$ research units, the allocation which is optimal in the nonstochastic case in which $\hat{\theta}$ is fixed at zero, is represented by $\frac{1}{2}[\mathit{Area}\ \mathit{VEF}-\mathit{Area}\ \mathit{AVD}+\mathit{Area}\ \mathit{FHG}-\mathit{Area}\ \mathit{HJK}]$. This expression could well be negative, even though this option avoids

---

[9] In accordance with the advice of a referee, the discussion below concentrates on a heuristic graphical exposition; a mathematical derivation of the expected net gain from using a patent, relative to contracting for $m_0$ units of research, is relegated to the Appendix. The expected gain from a prize is a special case of the gain from a patent, with $L$ and $\sigma_\zeta^2$ fixed at zero.

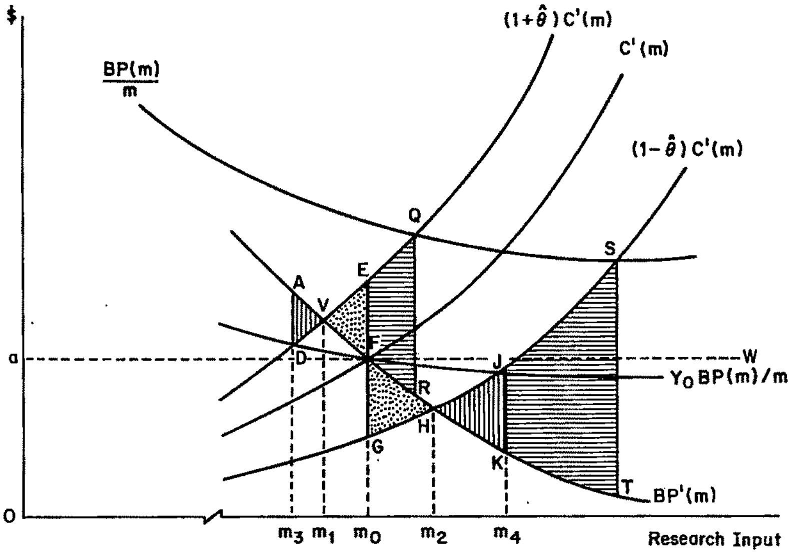


FIGURE 1. THE SOCIAL VALUE OF PRIZES WHEN RESEARCHERS HAVE EXCLUSIVE COST INFORMATION

the alternative welfare loss, represented by $\frac{1}{2}$[*Area EQRF* + *Area JSTK*], which would be incurred if a prize were offered equal to $B$, the social value of success.

Figure 1 can be used to examine an alternative to the prize option, namely a commitment by the planner to purchase units of research activity at a fixed price, announced at time 0, represented by the horizontal dashed line $aW$. (This option is equivalent to the "price" option for pollution control presented by Weitzman, 1974.) It is easy to show that the prize option is superior to this alternative; its advantage is that it offers an incentive with a desirable positive correlation with the equilibrium marginal productivity of research.

So far I have illustrated the gains from choosing a prize rather than direct contracting, for just one example. However, as shown in the Appendix, it is possible to derive more general results from this analysis, expressed in terms of elasticities, for cases in which the stochastic shifters $\zeta$ and $\theta$ have sufficiently small variance. For a prize, the derivation can be illustrated as follows.

An alternative measure of the expected net differential gain in Figure 1 is [*Area DEF* − *Area AFD* + *Area FJG* − *Area FJK*]. We can approximate these areas with the areas of triangles bounded by one straight line through point $E$, two through $F$, and one through $G$, with slopes equal to $C''(m_0)$, $BP''(m_0)$, $C''(m_0)$, $\partial[Y_0BP(m_0)/m_0]/\partial m$, and $C''(m_0)$, respectively.

Areas $DEF$ and $FHG$ have a vertical side of length $a\hat{\theta}$. Triangles approximating areas $AFD$ and $FJK$ have a vertical side of length $(\partial[Y_0BP(m_0)/m_0 - BP'(m_0)]/\partial m)\Delta m$, where $\Delta m$ approximates the distances $(m_0 - m_3)$, and $(m_4 - m_0)$,

$$\Delta m = a\hat{\theta} \Big/ \frac{\partial}{\partial m}\left[C'(m_0) - \frac{Y_0BP(m_0)}{m_0}\right].$$

Thus if $\hat{\theta}$ is sufficiently small, the expected net differential social gain from a prize relative to contracting for $m_0$ units of research, $G(prize)$, can be approximated as the expected difference between two triangles:

$$(5)\quad G(prize) = \frac{1}{2}E\left\{a\hat{\theta}\Delta m - \frac{\partial}{\partial m}\left[Y_0\frac{BP(m_0)}{m_0} - B\cdot P'(m_0)\right](\Delta m)^2\right\}.$$

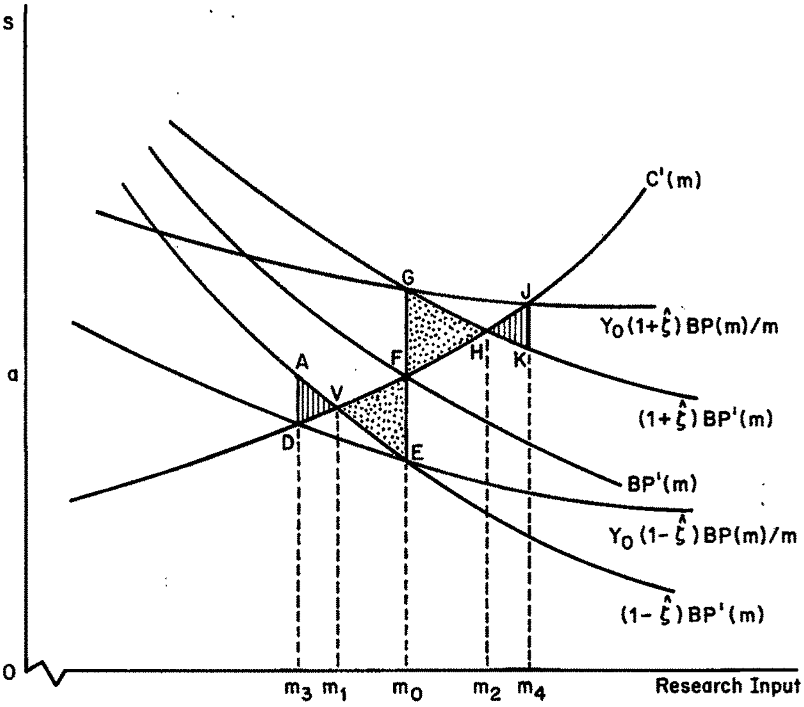


FIGURE 2. THE SOCIAL VALUE OF PATENTS WHEN RESEARCHERS HAVE EXCLUSIVE INFORMATION ABOUT THE BENEFITS OF SUCCESS

Substitute the expression for $\Delta m$ above, multiply above and below by $m_0/a$, and use equation (4) to obtain

$$(6)\quad G(prize)=\frac{am_0\sigma_\theta^2}{2[1/\eta-h(AP)]}\times\left\{1+\frac{h(MP)-h(AP)}{1/\eta-h(AP)}\right\},$$

where $\eta>0$ is the elasticity of supply of $m$ at $m_0$, $\eta(AP)=(m_0P'(m_0)/m_0-1)<0$ is the elasticity of the average probability of success, $P(m_0)/m_0$, and $h(MP)=m_0P''(m_0)/P'(m_0)<0$ is the elasticity of the marginal probability of success.

If the patent welfare cost $L$ equals zero, the gain from choosing a prize in the case illustrated in Figure 1 also equals the gain from choosing a patent. More generally, the difference between the social value of prizes shown in Figure 1 and the social evaluation of patents is the difference between two components. The first is the expected social value arising from the fact that patents, unlike prizes, give researchers an incentive to use exclusive information regarding benefits $B$, and the second is the patent deadweight loss $L$. The value of the first component is explored in Figure 2, in which $L$ and $\hat{\theta}$ are assumed to be fixed at zero. Benefits $B$ are subject to the random shifter $\tilde{\xi}$, which takes either of two values $\pm\hat{\xi}$, each with a probability of one-half. Either $m_3$ or $m_4$ units of research are produced. The expected net social benefit of patents is represented by $\frac{1}{2}[Area\ VEF-Area\ AVD+Area\ GHF-Area\ HJK]$, which is not necessarily positive. An alternative expression for this area is $\frac{1}{2}\{Area\ DEF-Area\ DAE+Area\ FJG-Area\ GJK\}$. These areas can be approximated by triangles bounded by two straight lines through $E$, one through $F$, and two through $G$, with slopes $\partial[Y_0BP(m_0)/m_0]/\partial m$, $BP''(m_0)$, $BP''(m_0)$, $\partial[Y_0BP(m_0)/m_0]/\partial m$, and $BP''(m_0)$, respectively.

Areas $DEF$ and $FJG$ in Figure 2 have a vertical side of length $a\hat{\xi}$. Triangles approximating areas $DAE$ and $GJK$ have a vertical side of length $(\partial[Y_0BP(m_0)/m_0 - BP'(m_0)]/\partial m)/\Delta m$, where $\Delta m$ approximates the distance $(m_0 - m_3)$, and $(m_4 - m_0)$:

$$\Delta m = a\hat{\xi} \Big/ \frac{\partial}{\partial m}\left[C'(m_0) - Y_0\frac{BP(m_0)}{m_0}\right].$$

Thus if $\hat{\xi}$ is sufficiently small, and independent of $\theta$, and the patent deadweight loss $L = 0$, the expected differential social gain from choosing a patent rather than a prize can be approximated as the expected difference between two triangles, that is,

$$G(patent)|_{L=0} - G(prize) = \frac{1}{2}E\left\{a\hat{\xi}\Delta m - \frac{\partial}{\partial m}\left[Y_0\frac{BP(m_0)}{m_0} - BP'(m_0)\right]\cdot(\Delta m)^2\right\}. \tag{7}$$

Substitution for $\Delta m$ in (7), use of (4), and multiplication above and below by $m_0/a$ yields the following approximation for the expected net gain illustrated in Figure 2:

$$G(patent)|_{L=0} - G(prize) = \frac{am_0\sigma_\xi^2}{2[1/\eta - h(AP)]} \times\left\{1 + \frac{h(MP) - h(AP)}{1/\eta - h(AP)}\right\}. \tag{8}$$

The expected deadweight loss from a patent scheme, ignored in Figure 2, is approximately,

$$Y_0LP(m_0) = m_0\frac{P'(m_0)}{P(m_0)}\cdot P(m_0)\cdot L = am_0\frac{L}{B},$$

since $BP'(m_0) = a$. Subtraction of this from the sum of equations (8) and (6) reproduces the approximation derived in the Appendix for the expected net social gain $G(patent)$ from using a patent, relative to opting for $m_0$ units of contractual research, when $Y$ is optimally adjusted by the administrator, given his information,

$$G(patent) = a\cdot m_0\left\{\frac{(\sigma_\xi^2 + \sigma_\theta^2)}{2[1/\eta - h(AP)]} \times\left[1 + \frac{h(MP) - h(AP)}{1/\eta - h(AP)}\right] - \frac{L}{B}\right\}. \tag{9}$$

This approximation, which eliminates terms of smaller order than $\sigma_\xi^2$ and $\sigma_\theta^2$, is appropriate not only for the two-point probability density functions in the examples illustrated in Figures 1 and 2, but also for any independent compact distributions of $\hat{\xi}$ and $\hat{\theta}$ with mean of zero, provided that the variances are sufficiently small. Under the same conditions, equation (6), which is a special case of (9), is an appropriate approximation to the expected social gain from using a prize incentive rather than directly contracting for $m_0$ units of research.

In the model of the previous section, patent and prize incentives, optimally adjusted *ex ante*, were unaffected by the common pool problem. But Figures 1 and 2, and expressions (6) and (9), show that this is not true when there is an informational asymmetry of the type which one might suppose would justify the more decentralized invention incentives. The common pool problem is then a real drawback of patents and prizes even if the limitation on the award is optimized *ex ante*. To show the extent of the efficiency losses caused by the common pool problem, let us now consider the expected differential net gains from the yardstick marginal patents and prizes discussed in Section I, which are unaffected by the common pool problem even if information is asymmetrically distributed. They thus are consistent with standard analyses of models with no informational asymmetry, in which the common pool problem (as well as any reason for choosing decentralized incentives) does not appear.

[3] *Marginal Patent.* Under the marginal patent, the administrator is again assumed to have the ability to optimally reduce expected

patent value to a fraction, denoted by $Z$, of its unrestricted level. In this case, the sole rationale for such a "patent life reduction" would be to reduce optimally the excess burden $L$, as in Nordhaus, who ignores the common pool problem. In this model, the equivalent of patent life limitation is to award the patent with conditional probability $Z$, given exactly one research project is successful. When $L$ equals zero, $Z$ obviously equals unity.

[4] *Marginal Prize*. Like the marginal patent, this prize is awarded only in the case where just one research effect is successful in making the particular discovery which is the objective of the incentive.

In the cases illustrated in Figures 1 and 2, the expected social gain from choosing a marginal patent relative to contracting for $m_0$ units of research is represented by $\frac{1}{2}$[*Area VEF* + *Area FHG*], which is nonnegative. The expected advantage of the marginal patent over the more feasible patent considered above, entirely attributable to the interaction of the common pool problem with the informational asymmetry, is represented by $\frac{1}{2}$[*Area AVD* + *Area HJK*] for both the alternate cases illustrated in Figures 1 and 2.

In either figure, the ratio of the triangular approximation for *Area VEF* to the approximations for *Area DEF* derived above is equal to the approximation to $m_0 - m_1$, divided by the approximation to $m_0 - m_3$ made above. This ratio is $[1/\eta - h(AP)]/[1/\eta - h(MP)]$. It is easy to infer from equation (9) and the discussion of its components[10] that

$$(10)\quad G(\mathit{marginal\ patent}) = \frac{am_0\left(\sigma_\zeta^2 + \sigma_\zeta^2\right)}{2(1/\eta - h(MP))} - am_0\frac{L}{B};$$

$$(11)\quad G(\mathit{marginal\ prize}) = \frac{am_0\sigma_\theta^2}{2(1/\eta - h(MP))}.$$

Subtraction of equation (9) from (10) yields:

$$(12)\quad G(\mathit{marginal\ patent}) - G(\mathit{patent}) = \frac{am_0\left(\sigma_\zeta^2 + \sigma_\zeta^2\right)[h(AP) - h(MP)]^2}{2[1/\eta - h(MP)][1/\eta - h(AP)]^2}.$$

Similarly

$$(13)\quad G(\mathit{marginal\ prize}) - G(\mathit{prize}) = \frac{am_0\sigma_\theta^2[h(AP) - h(MP)]^2}{2[1/\eta - h(MP)][1/\eta - h(AP)]^2}.$$

Given $1/\eta > 0 > h(MP)$, equations (12) and (13) are nonnegative.

The comparisons with the less practical marginal options have shown why the more feasible patents and prizes initially considered cannot motivate researchers holding relevant information unavailable to the public sector to allocate the expected social welfare-maximizing amount of resources to research. *This problem exists even if the researchers have all the available information and know all the relationships in the model, and the administrator optimally sets the patent life or prize value ex ante, given his information set.*

## IV. An Illustrative Example

To show how the parameters of the model interact to determine the appropriate choice of incentive when $\sigma_\theta^2$ and $\sigma_\zeta^2$ are sufficiently small, to show the cost of choosing a less effective incentive, and to allow checks on the validity of the approximations in Section III when these variances are finite, the following simple example is offered. Because marginal and average probabilities of success are relevant, it is necessary to choose a plausible functional form, rather than just a local approximation, for $P(m)$. In the model of Section II, there is a set of potential research avenues or "projects," each requiring one unit of research input and one period for completion, and each having the same probability of success in finding the single discrete invention goal. If it is further assumed that researchers can recognize which avenues are being pursued by competitors, have sufficient alternative avenues, and can avoid du-

---

[10] A derivation for $G$(*marginal patent*) in equation (11), similar to derivation of $G$(*patent*) in the Appendix, is available from the author.

TABLE 1—SIGNIFICANCE OF THE COMMON POOL PROBLEM: THE NET EXPECTED BENEFITS OF THE PRIZE OPTION RELATIVE TO THE YARDSTICK MARGINAL PRIZE[a]

| Probability that Success is Achieved[b] | Elasticity of Supply of Research | | | | | |
|---|---|---|---|---|---|---|
| | .1 | .5 | 1 | 2 | 10 | $\infty$ |
| .1 | 1.00 | 1.00 | 1.00 | 0.99 | 0.87 | c,d |
| .3 | 1.00 | 0.99 | 0.97 | 0.92 | 0.50 | c |
| .6 | 1.00 | 0.95 | 0.86 | 0.65 | c | c |
| .9 | 0.98 | 0.68 | 0.19 | c | c | c |

[a] Figures in the body of the table represent $G(prize)/G(marginal\ prize)$ from equations (6) and (11), assuming $\sigma_\theta^2 > 0$.

[b] $P(m) = 1-(1-\pi)^m$, evaluated at the nonstochastic optimum $m_0$, with $\pi = 0.01$; $P(m)$ is the probability that at least one of $m$ research projects succeeds.

[c] The decentralized options are dominated by a contract for $m_0$ units of research.

[d] Accurate only for very small values of $\sigma_\theta^2$.

plication of the research efforts of each other, the total and marginal probabilities of success (i.e., of at least one successful project) are

$$P(m) = 1-(1-\pi)^m; \tag{14}$$

$$P'(m) = -(1-\pi)^m \log(1-\pi), \tag{15}$$

where $\pi$ is the probability of success of any one project, assumed to be small. Then the significance of the common pool problem for sufficiently small $\sigma_\xi^2$, $\sigma_\theta^2$ is indicated by the ratio $G(prize)/G(marginal\ prize)$, reported in Table 1. The results are also bounds on the performance of the patent option relative to the yardstick marginal patent.

Note that toward the bottom right-hand corner of Table 1 the (optimized) prize option is seriously inferior, and for a considerable range of values the contractual allocation of $m_0$ by the administrator would be preferred, even though it wastes researchers' knowledge of $\theta$. The alternatives are nearly equivalent when the elasticity of supply and the equilibrium probability of success (and the level of research activity) are low. But as supply elasticity falls, the net benefits at stake also decline; incentives are important only if they elicit a nonnegligible response.[11]

If the yardstick marginal options are infeasible, what is the best incentive mechanism from the choice set which includes patents and prizes and direct government contracts? We know, from Table 1, that contracts can dominate prizes. The choice between prizes and patents depends on the tradeoff between the advantage of the latter mechanism, that it uses researchers' superior knowledge of the benefits of success, against its drawback, the deadweight loss associated with the patent.

The patent dominates the prize option if and only if

$$L/B\sigma_\xi^2 < \frac{1}{2[1/\eta - h(AP)]} \times \left[1 + \frac{h(MP)-h(AP)}{1/\eta - h(AP)}\right]. \tag{16}$$

Critical values of $L/B\sigma_\xi^2$ calculated using

---

[11] Given the number of approximations made in deriving them, the reader may wonder whether these results are accurate for finite $\sigma_\theta^2$. Consider then the case where the multiplicative shifter $\theta$ can take two values, $\pm 0.2$, each with probability 0.5, so that $\sigma_\theta^2 = 0.04$. Further, assume an industry marginal cost function

$$C = (1+\theta)[Am^{(1/\eta+1)}/(1/\eta+1)], A > 0, \eta > 0,$$

where $\eta$ is the (constant) supply elasticity. Then using equation (14) and searching for optimal reduction of the prize from $B$, one can confirm that the figures in the first five columns of the table are accurate within 0.01. The least accurate entry is in the upper right-hand corner; 0.87 should be 0.88 in this example. Paradoxically, linear supply curves (for which the approximation (3) is exact) yield greater inaccuracies in this example, for sufficiently low $P(m)$ and high $\eta$. At $P(m) = 0.1$, $\eta = 10$, research ceases when $\theta = +0.2$. Therefore $Y$ can be optimally adjusted for $\theta = -0.2$, rendering conventional prizes as efficient as marginal prizes, and superior to contracts. This problem diminishes as more points are added to the distribution of $\theta$, and as $\sigma_\theta^2$ decreases.

TABLE 2—THE CHOICE BETWEEN PATENTS AND PRIZES: CRITICAL VALUES[a] OF $L/B\sigma_\zeta^2$

| Probability that Success is Achieved[b] | Elasticity of Supply of Research | | | | | |
|---|---|---|---|---|---|---|
| | .1 | .5 | 1 | 2 | 10 | $\infty$ |
| .1 | 0.05 | 0.24 | 0.45 | 0.82 | 2.11 | c,d |
| .3 | 0.05 | 0.21 | 0.36 | 0.56 | 0.55 | c |
| .6 | 0.05 | 0.16 | 0.22 | 0.23 | c | c |
| .9 | 0.04 | 0.08 | 0.03 | c | c | c |

[a] When $L/B\sigma_\zeta^2$ is greater (less) than the number in the text, prizes (patents) are the optimal choice.

[b] $P(m) = 1 - (1 - \pi)^m$, evaluated at the nonstochastic optimum $m_0$, with $\pi = 0.01$.

[c] Patents and prizes are dominated by a contract for $m_0$ units of research.

[d] Accurate only for very small values of $\sigma_\theta^2, \sigma_\theta^2$.

(16) are presented in Table 2. Patents are most likely to be preferable to prizes when the overall success probability is low, and the elasticity of supply of research is moderately high. When probability of success and/or research supply elasticity is high, contracts can dominate patents.[12]

To help the reader judge whether the entries in Table 2 are likely to cover plausible magnitudes of $L/B\sigma_\zeta^2$, I offer one specific example. A standard error of 0.25 on the multiplicative shifter $\zeta$ would seem to bestow on researchers an ample advantage in their knowledge of the value of the invention.[13] Nordhaus (p. 81) estimates a typical cost reduction from a process invention at around 5 percent, and chooses 1.0 as an average value of the elasticity of product demand. These values imply that $L/B$ (the ratio of the welfare cost of the patent to the sum of the patent revenue and the welfare cost, discussed in fn. 7), is approximately 0.025.

These choices imply that $L/B\sigma_\zeta^2$ equals 0.4, which lies in the intermediate range of the values in Table 2. Note that a higher value of $\sigma_\zeta^2$, or a lower value of $L/B$, would not prevent domination by contracts in the lower right-hand corner of the table. Using the probability of success function (13), the choice between patents, prizes, and contracts, for $L/B\sigma_\zeta^2 = 0.4$, and $\sigma_\theta^2$ positive and small, is illustrated in Figure 3. As indicated, patents are optimal only in the shaded area toward the bottom of the figure. The argument of Machlup and Plant that an inelastic supply of inventors implies that a patent system is undesirable is supported by this analysis. Patents are most attractive at supply elasticities between 2.0 and 4.0, and especially for long shot research areas with low aggregate probability of success; patents are never the best choice if the aggregate probability of success is over 0.45. Prizes are best for research areas with intermediate success probabilities, and for all areas where supply is inelastic, as is likely if there are essential research-specific inputs, unless the overall research venture is almost certain to succeed.

When supply is elastic, contracts are the best choice for the considerable range of values covered by the upper shaded area, and they dominate increasingly as the overall success prospects become more certain or as supply elasticity increases. Thus contracts are best when the research process is most like activities routinely undertaken in other sectors of the economy, where sectoral input supplies are elastic and output is fairly predictable. Paradoxically, the more decentralized incentives are less attractive when they

---

[12] An accuracy check using a constant elasticity supply function, and with $\zeta = \pm 0.2$, each with probability 0.5, shows the figures in the first five columns accurate within 10 percent, that is, about as accurate as Table 1, given the difference in denominator. Once again, the caveat for linear supply with high $\eta$ and low $P(m)$ applies.

[13] Note that the coefficient of variation of $B$ could be very much greater than $\sigma_\zeta$ if, as one might expect, there are other random shifters, independent of $\zeta$, unobserved by all parties. Only the *relative* difference in information is measure by $\zeta$.

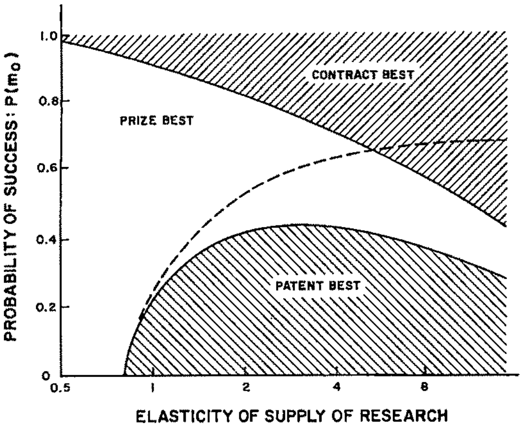


FIGURE 3. THE OPTIMAL CHOICE OF INVENTION INCENTIVE ($L/B\sigma_\xi^2 = 0.4$)

elicit a highly sensitive economic response! This result is due to the common pool problem; if it did not exist, or if marginal patents and prizes were feasible, patents would be the best choice in Figure 3 below the dashed line, and prizes would be best anywhere above that line.

## V. Generalization and Qualifications of the Results

At this point it is appropriate to consider how the choices between incentives change as the assumptions adopted above are modified. First consider the informational assumptions. It can be shown that the relative merit of each incentive does not change if the administrator has exclusive information about costs or returns which is not directly communicated to researchers. By optimally modifying the value of the prize or patent, the administrator can make researchers behave as if they share his knowledge of these random elements.[14] It is only exclusively *private* information which affects the choice between incentives. Furthermore if, instead of shifting $C(m)$, the stochastic term $\theta$ shifted $P(m)$ in an analogous fashion, the results derived above would be unchanged. Exclusively private information about the probability function, like exclusively private information about the cost function, increases the value of patents and prizes equally.

The special advantage of patents arises only from *ex ante* researcher information relating to the value of the invention. If, on the other hand, the administrator cannot *ex post* costlessly monitor the quality-adjusted effort of researchers (successful or not), the two more decentralized options should have a further advantage over contracts by allowing self-selection and self-supervision of researchers. In fact, a major real world disadvantage of contracts, ignored here, is the X-inefficiency which can be attributed to the lack of incentives for cost control by researchers in situations where external monitoring is difficult. (However, Yu, pp. 220–21, argues that enforceability problems of contracts have been exaggerated.) An extension of the current model could include differential monitoring costs for each incentive, and model X-inefficiency as a shift in the supply of research. An even more serious problem of moral hazard could arise under patents or prizes, if the patent life or amount of the award were modified *ex post* as a function of input costs, since cost guidelines cannot be as rigidly stipulated as under a contract if the greater private initiative under these options is to be given full rein. There is an additional moral hazard problem from the viewpoint of researchers, in that the government may understate its *ex post* evaluation if it wishes to minimize expenditures, and is not greatly concerned with the effects of such action on the reputation of future governments. For these reasons, and because, as Kitch has emphasized, even the current fixed-life patent system incurs a great burden of litigation, I have not considered awards whose terms are adjusted *ex post*. In any case, where such awards are feasible, patents are inferior to direct grants or prizes because of the presumed higher excess burden of patents. This may explain why we do not observe patents with lives determined *ex post*, whereas, as one referee has noted, most historical examples of prizes were determined *ex post*, though two very important inventions of the modern era, the chronometer and canning, were awarded predetermined prizes for prespecified achievements, like the prizes considered here.[15]

---

[14] An analysis of these cases is available from the author.

Finally, the extension to an intertemporal framework of the type pioneered by Barzel appears to be quite feasible for the analysis of patent and prize options presented here. However, as mentioned above, anticipation of severe practical difficulties with implementation of the marginal patent and marginal prize options in an intertemporal context is reflected in the fact that they are treated here as convenient yardsticks rather than as realistic alternatives.

## VI. Conclusions

In the model presented here, the peculiar informational advantage of patents over prizes is that they incorporate in the allocative process researchers' information about the value of successful inventions, if such information is unavailable to the public research administrator. Both patents and prizes utilize private information about research costs, or the probability of success. However, information held exclusively by the public administration has no effect on the choice between all three options, optimally implemented.

The informational advantages of the more decentralized research incentives do not necessarily translate into an allocative advantage. Though the existence of exclusively private information is necessary to render patents and prizes superior to contracts, it is by no means sufficient, because of the common pool problem. Even if the value of patents and prizes can be adjusted by a fraction that is optimal *ex ante*, such reduction does not eliminate the effects of this problem in prizes and patents, as it does in previous analytical models where all agents have identical information. When this problem is recognized, the paradoxical result can emerge that contracts, the centralized alternative, are more likely to be the best choice if researchers are highly responsive to incentives than if they are less responsive! If contracts dominate prizes, then patents are the worst choice of all in the model presented here. If contracts are inferior, the incentive choice, for a patentable invention, rests on the tradeoff between the excess burden of the patent system and its informational advantage over prizes.

More specific conclusions about the choice of incentive depend on the nature of the search process, and just one of many plausible models is analyzed in Section IV. Future studies could investigate alternative search processes and relaxation of the assumptions of a discrete goal of research, fully competitive researcher supply, risk neutrality, and the homogeneity of researcher attributes. As a prelude to such further investigation, this paper suggests that the range of situations in which a practical patent system dominates other feasible alternatives may be narrower than is commonly believed, while the relative advantages of contractual research and of prizes may well have been undervalued.

## APPENDIX: DERIVATION OF THE EXPECTED NET SOCIAL GAIN FROM CHOOSING THE PATENT OPTION

If the patent instrument with conditional award probability $Y$ is adopted in the model of Section II in which multiplicative stochastic shifters $\zeta$ and $\theta$ are introduced, the difference in the incentive from equation (3), which refers to the model of Section I, is

$$\left[YB(1+\zeta)(1-L/B)Pm^{-1}-Y_0B\bar{P}m_0^{-1}\right].$$

Denote a first-order approximation to this expression, evaluated at $m=m_0$, $Y=Y_0$, $\zeta=L=0$, by $\Delta I$,

$$\text{(A1)}\quad \Delta I=Y_0\frac{B}{m_0}P'(m_0)(m-m_0)+Y_0\frac{B\bar{P}}{m_0}\left(\zeta-\frac{L}{B}\right)+\frac{B\bar{P}}{m_0}(Y-Y_0).$$

---

[15] Awards of £10,000, £15,000, and £20,000 were offered by the British Board of Longitude in 1713 for a chronometer which measured longitude to within 60, 40, and 30 minutes, respectively. John Harrison claimed the £20,000 reward in 1762, and full payment was completed by 1773 (*Encyclopaedia Britannica*, 1929, Vol. 11, p. 220). In 1795 a prize of 12,000 francs was offered by Napoleon's Society for the Encouragement of Industry for a method of food preservation useable by the military forces. It was awarded in 1810 to Nicolas Appert for a method of food canning using heat treatment of food in sealed champagne bottles (William Harris and Judith Levey, 1975, p. 127; James Burke, 1978, p. 234).

Research firms' responses, which determine $m$, are based on their knowledge of the stochastic terms and the announced value of $Y$. In equilibrium, the number of active firms, each producing one unit of research, is

$$\text{(A2)} \qquad m = m(\zeta, \theta, Y).$$

Relative to the equilibrium of Section I, the equilibrium change in its marginal and average cost $c_1(1, m, \theta)$ perceived by any individual research firm equals the change in industry marginal cost $C_1(m, \theta)$. A first-order approximation to this expression, evaluated at $m = m_0, Y = Y_0, \theta = 0$, is

$$\text{(A3)} \qquad \Delta C_1 = a\theta + b(m - m_0),$$

where $a \equiv C'(m_0)$, $b \equiv C''(m_0)$.

Definition of $m_0$ as the equilibrium level of $m$ when $L$, $\zeta$ and $\theta$ are all fixed at zero implies that

$$\text{(A4)} \qquad a \equiv B \cdot P'(m_0).$$

Equating (A1) to (A3) and using (A2),

$$\text{(A5)} \quad m(\zeta, \theta, Y) - m_0 = \left[a\theta - \frac{1}{m_0}\left(Y_0 B\bar{P} + Y_0 \bar{P} L - B\bar{P}(Y - Y_0)\right)\right] \Big/ W,$$

where

$$\text{(A6)} \qquad W = Y_0 \frac{a}{m_0} - Y_0 \frac{B\bar{P}}{m_0^2} - b.$$

The administrator sets the fraction $Y$ so as to maximize expected net social benefits, given that the patent system is chosen. Throughout the analysis that follows the second-order conditions of maximization are assumed to be satisfied, and corner solutions are ruled out. Under these conditions, the first-order condition for maximization of net social benefits expected by the administrator is

$$\text{(A7)} \quad E_G\{(1+\zeta)(1 - L/B) \times [a + Bf(m(\zeta, \theta, Y) - m_0)] - (1+\theta)[a - b(m(\zeta, \theta, Y) - m_0)]\} = 0$$

where $E_G$ denotes the expectation of the government administrator conditional on his information in time 0, and $f \equiv P''(m_0)$. Substituting (A5) in (A7) using (A4),

$$\text{(A8)} \quad \frac{1}{m_0} \cdot B\bar{P}(Y - Y_0) = -\frac{Wa\frac{L}{B} - \sigma_\zeta^2 Y_0 \bar{P} B^2 f m_0^{-1} - ab\sigma_\theta^2}{(B-L)f - b} + \frac{Y_0 \bar{P} L}{m_0}.$$

Substituting (A8) in (A5) and using equation (4), and discarding terms of smaller order than $\sigma_\zeta^2$ and $\sigma_0^2$,

$$\text{(A9)} \quad m(\zeta, \theta, Y) - m_0 = a \cdot W^{-1}\left[\zeta - \theta + \frac{\sigma_\zeta^2 Bf + b\sigma_\theta^2}{(B-L)f - b}\right] + \frac{a\frac{L}{B}}{(B-L)f - b}.$$

The expected cost of the patent distortion is $E_G[P \cdot Y \cdot L]$. The net differential social gain expected by the administrator from choosing the limited patent instrument in time 0 rather than directly contracting for research, $G(\mathit{patent})$ is

$$\text{(A10)} \quad G(\mathit{patent}) = E_G\Big[(1+\zeta)\Big[a(m(\zeta, \theta, Y) - m_0) + \frac{Bf}{2}(m(\zeta, \theta, Y) - m_0)^2\Big] - (1+\theta)\Big[a(m(\zeta, \theta, Y) - m_0) - \frac{b}{2}(m(\zeta, \theta, Y) - m_0)^2\Big] - Y \cdot L \cdot P\Big].$$

Substituting (A9) and (A6) in (A10) and

discarding terms of smaller order than $\sigma_\xi^2$ and $\sigma_\theta^2$,

$$\text{(A11)}\quad G(patent) = am_0\left\{\left[\frac{\sigma_\xi^2+\sigma_\theta^2}{\left(\frac{1}{\eta}-h(AP)\right)}\right]\times\left[1-\frac{\frac{1}{\eta}-h(MP)}{2\left(\frac{1}{\eta}-h(AP)\right)}\right]-\frac{L}{B}\right\}$$

where $\eta \equiv a/bm_0 \geqslant 0$ is the elasticity of supply of research activity at $m_0$, and $h(MP)=bfm_0/a$ and $h(AP)=am_0/\bar{P}-1$ are the elasticities of the marginal and average probabilities of success, respectively. Rearranged, (A11) becomes equation (9) in the text.

## REFERENCES

**Arrow, Kenneth J.**, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, New York: Arno Press, 1962.

**Barzel, Yoram**, "Optimal Timing of Innovations," *Review of Economics and Statistics*, August 1968, *50*, 348–55.

**Bowman, Ward S., Jr.**, *Patent and Antitrust Law*, Chicago: University of Chicago Press, 1973.

**Burke, James**, *Connections*, London: MacMillan, 1978.

**Cheung, Steven N. S.**, "Property Rights in Trade Secrets," *Economic Inquiry*, January 1982, *20*, 40–52.

**Clarke, Edward H.**, "Multipart Pricing of Public Goods," *Public Choice*, Fall 1971, *11*, 17–33.

**Dasgupta, Partha and Stiglitz, Joseph**, (1980a) "Uncertainty, Industrial Structure and the Speed of *R&D*," *Bell Journal of Economics*, Spring 1980, *11*, 1–28.

______, (1980b) "Industrial Structure and the Nature of Innovative Activity," *Economic Journal*, June 1980, *90*, 266–93.

**Gilbert, Richard J. and Newbery, David M. G.**, "Preemptive Patenting and the Persistance of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

**Gordon, H. Scott**, "The Economic Theory of a Common-Property Resource: The Fishery," *Journal of Political Economy*, April 1954, *62*, 124–42.

**Harris, William H. and Levey, Judith S.**, *The New Columbia Encyclopedia*, New York: Columbia University Press, 1975.

**Kamien, Morton I.**, "Patent Life and *R&D* Rivalry," *American Economic Review*, March 1974, *64*, 183–87.

______ **and Schwartz, Nancy L.**, "Market Structure and Innovation: A Survey," *Journal of Economic Literature*, March 1975, *13*, 1–37.

**Kitch, Edmund W.**, "The Nature and Function of the Patent System," *Journal of Law and Economics*, October 1977, *20*, 265–90.

**Kitti, Carole**, "Patent Policy and the Optimal Timing of Innovations," unpublished doctoral dissertation, University of Chicago, 1973.

**Laffont, Jean Jacques**, "More on Prices vs. Quantities," *Review of Economic Studies*, February 1977, *44*, 177–82.

**Lee, Tom and Wilde, Louis L.**, "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, *94*, 429–36.

**Loury, Glenn C.**, "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93*, 395–410.

**Machlup, Fritz**, *An Economic Review of the Patent System*, Study No. 15 of the Subcommittee on Patents, Trademarks, and Copyrights of the Committee on the Judiciary, U.S. Senate, 85th Congress, 2nd Session, Washington, 1958.

**Malcomson, James M.**, "Prices vs. Quantities: A Critical Note on the Use of Approximation," *Review of Economic Studies*, February 1978, *45*, 203–07.

**Nelson, Richard R.**, "The Economics of Invention: A Survey of the Literature," *Journal of Business*, April 1959, *32*, 101–27.

______, "The Role of Knowledge in R&D Efficiency," *Quarterly Journal of Economics*, August 1982, *97*, 453–70.

**Nordhaus, William D.**, *Invention, Growth and Welfare*, Cambridge: MIT Press, 1969.

**Plant, Arnold,** "The Economic Theory Concerning Patents and Inventions," *Economica*, February 1934, *1*, 30–51.

**Scherer, F. M.,** "Firm Size, Market Structure, Opportunity, and the Output of Patented Invention," *American Economic Review*, December 1965, *55*, 1097–124.

______, "Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation," *American Economic Review*, June 1972, *62*, 422–27.

______, *The Economic Effects of Compulsory Patent Licensing*," Graduate School of Business Administration, Center for the Study of Financial Institutions, Monograph 1977–2, New York University, 1977.

**Stiglitz, Joseph E.,** "Discussion," *American Economic Review Proceedings*, May 1969, *59*, 46–49.

**Taylor, Charles T. and Silberston, Z. A.,** *The Economic Impact of the Patent System*, Cambridge: Cambridge University Press, 1973.

**Usher, Dan,** "The Welfare Economics of Invention," *Economica*, August 1964, *31*, 279–87.

**Weitzman, Martin L.,** "Prices vs. Quantities," *Review of Economic Studies*, October 1974, *41*, 477–91.

______, "Reply to 'Prices vs. quantities: A Critical Note on the Use of Approximations' by James M. Malcomson," *Review of Economic Studies*, February 1978, *45*, 208–09.

**Wright, Brian D.,** "Is Protection of the Secrecy of Research Socially Desirable?," paper presented at the Annual Meetings of the Eastern Economics Association, Philadelphia, April 10, 1981.

**Willig, Robert D.,** "Consumer's Surplus Without Apology," *American Economic Review*, September 1976, *66*, 589–97.

**Yu, Ben T.,** "Potential Competition and Contracting in Innovation," *Journal of Law and Economics*, October 1981, *24*, 215–38.

***Encyclopaedia Britannica,*** 14th Edition, New York, 1929.

**National Bureau of Economic Research,** *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Universities-National Bureau Conference Series, No. 13, New York: Arno Press, 1962.

# Structural/Frictional vs. Deficient Demand Unemployment: Some New Evidence

*By* KATHARINE G. ABRAHAM*

Macroeconomic policymakers devote a substantial fraction of their effort to combating unemployment. One fact that has perhaps not been sufficiently recognized is that the magnitude of our country's job vacancy rate greatly conditions the optimal strategy in this arena. The potential efficacy of microeconomic labor market policies designed to facilitate matching people with already existing jobs versus that of aggregate policies designed to create new jobs depends critically on the extent to which unemployment is attributable to frictional or structural problems as opposed to deficient demand; this can only be gauged with information both on those seeking work and on the availability of unfilled jobs.

Except during periods of deep recession such as were experienced in 1971–72 and 1974–75, there have always been those who asserted that the number of vacant jobs was large relative to the number of persons seeking work. During the middle 1960's, Arthur Burns apparently believed it to be possible that the number of vacant jobs actually exceeded the number of unemployed persons. A *Wall Street Journal* editorial which appeared late in that decade lamented that there had been little official recognition of the fact that a large part of the problem is simply matching jobless men with available jobs. A similar point of view is implicit in another *Wall Street Journal* assessment of the then-current unemployment problem written in June 1973 which concluded that "most [of the 4,300,000 unemployed] were probably out of work either by their own choosing or for reasons that have little to do with total economic growth." Spokesmen for the Reagan Administration have repeatedly stressed the availability of jobs, even as the unemployment rate has climbed to a new postwar high. President Reagan commented at a March 1981 meeting with members of Congress that his newspapers contained many pages of help wanted advertising and suggested that the unemployment problem might reflect programs which encourage people to delay taking jobs rather than a scarcity of employment opportunities; at a January 1983 press conference, he argued that the help wanted advertisements revealed employers to be "desperate" for skilled workers, so that training programs might offer a reasonable response to the unemployment situation.[1] Whether the beliefs which underlie such statements are in fact correctly held has extremely important implications for the selection of an unemployment policy. This article assesses their validity.

Section I briefly discusses efforts made during the 1960's and 1970's to collect information on the number of unfilled job openings in the United States and Canada, then brings these vacancy data together with appropriate unemployment data. Even after very generous allowances are made for sources of downward bias in the vacancy rates, and the unemployment rates have been redefined to exclude persons already attached to an employer (those on temporary layoff and those due to start a job within 30 days), comparison of the two rates suggests

---

*Assistant Professor of Industrial Relations, Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139. This paper has benefited from discussions with Mary Corcoran, David Gower, Paul Harrington, Lois Plunkert and Eric Van der Wolt. Martin Baily, Charles Brown, Stanley Fischer, Richard Layard, and the participants in seminars at the University of Chicago, Harvard University, and the Massachusetts Institute of Technology also offered helpful comments. I owe particular thanks to James Medoff for his many constructive suggestions. This work was supported by the Department of Labor under Contract No. J-9-M-9-0220.

[1] See *The Wall Street Journal* (1965, 1969, 1973) and *The New York Times* (1981, 1983).

that the number of persons seeking work has typically been much larger than the number of vacant jobs. Section II takes an alternative approach to estimating the relationship between the number of persons seeking work and the availability of unfilled job openings. Information on job vacancy durations is combined with information on the new hire rate to produce independent estimates of the vacancy rate, which are then compared with the relevant unemployment rates. It is reassuring that the conclusions reached via this alternative route to making unemployment versus vacancy comparisons are highly consistent with those reached in Section I. The final section offers a summary of the paper's central findings and some concluding thoughts.

## I. Estimates of the Unemployment-to-Vacancy Ratio

Employers in the United States and Canada have provided usable job vacancy data in connection with six pilot projects and longer-term survey efforts since the mid-1960's. These sources of vacancy data can be used to estimate the unemployment-to-vacancy ratio associated with different unemployment rate levels. However, the development of meaningful figures requires both that the published vacancy numbers be adjusted for possible downward bias (some available jobs may not be counted) and that the unemployment numbers they are compared with be adjusted for possible upward bias (some of those counted as unemployed may not truly want work).

### A. *Vacancy Rate Data*

What do we mean by a job vacancy? At a conceptual level, a job vacancy represents unsatisfied labor demand. The wage rate obviously plays an important role here. If wages were low enough, employers would want to hire an indefinitely large number of workers. Employers who are asked how many jobs they have vacant will reply based on their beliefs about the wages they will have to pay. Implicitly, then, any job vacancy rate will reflect the number of currently unfilled job openings at what employers perceive to be the market wage rate. The vacancy rate is thus analogous to the unemployment rate, insofar as any individual's decision to be in or out of the labor force will reflect his or her wage expectations.

At an operational level, the measurement of job vacancies requires that employers be given clear instructions regarding what should and what should not be reported on the survey forms. All of the data collection efforts discussed in this paper defined a job vacancy in almost identical fashion:[2]

> *Job vacancies* are defined as current, unfilled job openings in your establishment which are immediately available for occupancy by workers from outside your firm and for which your firm is actively seeking such workers. Included are full-time, part-time, permanent, temporary, seasonal and short-term job openings. [p. 176]

Pilot job vacancy projects conducted by the Bureau of Employment Security (BES) and the Bureau of Labor Statistics (BLS) between the end of 1964 and April 1966 produced the earliest usable job vacancy information for the United States. Data were collected in fourteen or fifteen cities accounting for approximately 25 percent of the nation's total employment on each of three survey occasions; only agriculture, private households, and the military were not covered.[3] Based on this pilot project experience, it was decided that collection of vacancy data on an ongoing basis was feasible. From April 1969 through December 1973, job vacancy information for the U.S. manufacturing sector was collected by BLS under their *Job Openings and Labor Turnover Survey* (*JOLTS*) program. Only members of the panel of establishments already participating in the labor turnover data collection effort were solicited for participation in the *JOLTS* program; the original intention was to expand the panel to include a substantial number of nonmanufacturing establishments, but this goal was never attained.[4] When BLS discontinued funding for the job openings part of the *JOLTS* program, the states of Minnesota and Wisconsin chose to continue collecting job openings data on their own. These states were relatively successful in expanding coverage to nonmanufacturing sectors. Minnesota data for all sectors except agriculture, private households, railroads, construction, and the military are available from 1972 forward; Wisconsin data were produced for all sectors except agriculture, private households, and the military beginning in 1976.[5] Interest in job vacancy data collection revived during the late 1970's. The Congress appropriated money for a pilot job vacancy survey which ultimately yielded information for each of four cooperating states (Florida, Massachusetts, Texas, and Utah). This pilot study covered all major private nonagricultural sectors, excluding only private households, and produced job vacancy data once every three months from March 1979 through June 1980, when these experiments were discontinued.[6]

---

[2] This wording came from Subcommittee on Economic Statistics (1966).

[3] Paul Harrington (undated) discusses these early pilot vacancy projects in considerable detail. Hearings on the pilot projects were held before the Subcommittee on Economic Statistics of the Joint Economic Committee of the U.S. Congress in May 1966; the published report on these hearings contains copies of the various survey forms that BES experimented with in addition to testimony from a number of individuals.

In Canada, pilot job vacancy data collection efforts during the middle 1960's were followed by the launching of the full-scale *Job Vacancy Survey* (*JVS*) in 1970. A large random sample of establishments was selected for participation in the *JVS*. Because of concern that no single respondent in a large establishment would be able to give an accurate report on all current job openings, large establishments were broken into subunits on the basis of where hiring points were located and these subunits became the sampled entities. The survey was conducted in two phases: a mail phase; and a follow-up interview phase. The results of the interview phase were used to correct for systematic over- or underreporting on mail forms and for nonresponse bias. The *JVS* produced quarterly data for the entire Canadian nonagricultural economy, excluding only fishing and trapping, private households, and the military, from the start of 1971 through the end of 1978, when the program was discontinued.[7]

Data on job openings which cover a variety of time periods and geographic areas are thus available: 1) for selected U.S. cities, we have data for each of three dates during the mid-1960's; 2) data for the U.S. manufacturing sector are available for the period from April 1969 through December 1973; 3) nonagricultural vacancy rates have been produced in Minnesota since 1972; 4) nonagricultural vacancy rates have been produced in Wisconsin since 1976; 5) the recent BLS pilot program yielded 1979 and 1980 data covering four states; and 6) Canadian *JVS* data were generated from the start of 1971 through the end of 1978.

### B. *Correcting for Downward Bias in the Reported Vacancy Rate Data*

The primary potential sources of downward bias in the reported vacancy rates are listed in Table 1, along with a description of attempts to assess the likely importance of each bias and my reading of the evidence on each point.[8] The table underscores a number of important facts about efforts to measure the overall job vacancy rate. The most serious problem with available vacancy rate estimates seems to be that the firms which supply vacancy information are not representative of the universe of potential respondents. Large units tend to have both higher survey response rates and lower vacancy rates than small units; if an unstratified vacancy rate estimator is employed (as in the *JOLTS* program), this combination leads to downward bias in the estimated vacancy rate. Furthermore, even among units of similar size, nonrespondents appear to have substantially higher vacancy rates than respondents. While dissimilarities between vacancy survey respondents and nonrespondents may indeed cause substantial bias in reported vacancy rate estimates, other often-cited potential problems with existing vacancy data appear to be less consequential than has been believed. A particular concern has been that the vacancy concept used in collecting existing data might exclude large numbers of potential employment opportunities. The available evidence implies that "discouraged vacancies" (jobs which employers would like to fill but have given up recruiting for) and hires of walk-in candidates without prior recruiting are in fact relatively rare.

---

[4] A description of the *JOLTS* program, including a copy of the survey form used, can be found in U.S. Department of Labor (1971, pp. 35–42). Paul Armknecht (1974) summarizes the results of the job openings part of the survey.

[5] Harrington (1978) discusses the Minnesota and Wisconsin *JOLTS* programs. Conversations with Richard Johnson of the Minnesota Department of Economic Security and Kenneth D. Siemers of the Wisconsin Department of Industry, Labor and Human Relations yielded additional information concerning those states' vacancy data collection efforts.

[6] See U.S. Department of Labor (1981a) for a detailed discussion of the BLS 1979–80 vacancy pilot project experience.

[7] For a review of the Canadian *JVS* experience, see Harrington and William Milligan, undated. Copies of the various survey forms employed can be found in Statistics Canada (1972).

[8] Two concerns omitted from the enumeration in Table 1 are that inadequate record-keeping practices and/or the existence of multiple hiring points inside the firm might preclude the collection of accurate job vacancy statistics. The available evidence on these points, including the stated judgements of analysts who were

Combining the effects of all the possible sources of downward bias in Table 1 produced approximately the following aggregate correction factors for the various vacancy surveys: BES/BLS 1964–66 pilot projects, 1.70; U.S. *JOLTS* program, 2.11; Minnesota *JOLTS* program, 2.17; Wisconsin *JOLTS* program, 2.14; BLS 1979–80 pilot projects, 1.73; and Canadian *JVS* program, 1.33.[9]

Applying these correction factors to the raw vacancy rates from each source very likely leads to overstating the true level of job availability in the sectors covered by the various surveys. For one thing, fairly generous allowance was made for the impact of each possible source of downward bias in the reported vacancy estimates. In addition, no consideration was given to possible sources of *overstatement* in the raw figures. For example, some jobs with future starting dates were likely counted as currently available and some of the positions listed as vacant were likely offered at a below-market wage.[10]

Another important concern is whether the vacancy rate in the sectors covered by the various surveys is comparable to the rate in the excluded sectors. If there were reason to believe that the rate in the uncovered sectors exceeded the rate in the covered sectors, then even the corrected-for-downward-bias figures just discussed might understate the true *overall* vacancy rate. Assume that the inflated vacancy rate estimates just created accurately reflect the availability of jobs in the covered sectors:

$$VR_c = [V_c/(V_c + E_c)] \times 100, \tag{1}$$

where $VR_c$ represents the vacancy rate in the covered sectors, $V_c$ represents the number of jobs vacant in the covered sectors and $E_c$ represents employment in those same sectors. For the purpose of what follows, the relevant vacancy rate is equal to

$$VR = [V/(V + E)] \times 100, \tag{2}$$

where $VR$ represents the economywide vacancy rate, $V$ represents the total number of vacancies in the economy, and $E$ represents total employment in the economy. As discussed above, all of the vacancy data collection efforts other than the national *JOLTS*

---

intimately involved with various vacancy survey efforts, lead me to conclude that these two factors should not cause serious reporting problems. For a fuller discussion of these issues, see my 1982 dissertation (pp. 33–35).

[9] The formula used for calculating the aggregate correction factors was $CF_j = \Pi_i(1 + p_{ij})$, where $CF_j$ is the aggregate correction factor for survey $j$ and $p_{ij}$ is the proportional understatement which bias $i$ might reasonably have caused in the survey $j$ vacancy estimates (from Table 1). This calculation assumes that the various sources of bias are independent.

[10] My understanding is that the AFL-CIO has opposed the collection of aggregate vacancy data. This opposition may stem from fear that a large number of jobs paying below-market wages would be counted, which might lead policymakers to place too little emphasis on job creation; see, for example, Marvin Friedman (1966).

TABLE 1—POSSIBLE SOURCES OF DOWNWARD BIAS IN AVAILABLE VACANCY RATE ESTIMATES

| Source of Downward Bias in Reported Vacancy Rate and Corrections Applied | Evidence on Magnitude of Downward Bias |
|---|---|
| 1) Mail solicitations may not elicit as careful a response as interview solicitations. Firms may omit some vacancies from a mail form because they fear employment service referrals, because they don't understand that all occupations are supposed to be covered, or because giving an incomplete listing is easier than filling out the form correctly.<br>All the U.S. vacancy estimates were inflated by 10 percent to correct for this problem. The *JVS* estimates are already corrected for interview vs. mail bias. | Response analysis interviews were conducted in several cities following the second and third rounds of the mid-1960's BES/BLS pilots. The vacancy rate estimates which incorporated this interview information were 3–4 percent higher than those based on mail responses alone. (Harrington, undated, pp. 53 and 85.)<br>Response analysis interviews were conducted at a total of 80 establishments following the first round of 1979–80 BLS pilots. Comparisons of the vacancy rates obtained from the interviews against those obtained from the original mail forms were reported only for selected subsets of the sample. The interview-based vacancy rates ranged from 10 percent lower to 32 percent higher than the mail-based vacancy rates. (U.S. Department of Labor, 1981a, ch. 3.)<br>A 1966 Canadian pilot project produced vacancy estimates based on interviews with 162 firms in Windsor and Cornwall that were 70 percent higher than those based on mail responses from the same firms. (Dominion Bureau of Statistics, Part III, 1966.)<br>Calculations based on the very large amount of *JVS* microdata from the six survey occasions during February, March, and April, 1976, showed that correcting for net differences in interview as compared to mail responses increased the number of vacancies by an average of less than 10 percent. |
| 2) Nonresponse bias may be a problem if cooperating firms are not representative of the universe.<br>It was assumed that nonrespondents had a vacancy rate 35 percent higher than respondents. This led to inflating the BES/BLS 1964–66 pilot project vacancy estimates by 11 percent; the U.S. *JOLTS* estimates by 19 percent; the Minnesota *JOLTS* estimates by 23 percent; the Wisconsin *JOLTS* estimates by 21 percent; and the BLS 1979–80 pilot project estimates by 12 percent. The variation in the correction factors reflects variation in response rates, with larger corrections made where response rates were lower. The *JVS* estimates are already corrected for nonresponse bias. | The response analysis interviews which followed the first round of mid-1960's BES/BLS pilots produced a nonrespondent vacancy rate 35 percent higher than the respondent vacancy rate. (Harrington, undated, p. 53.)<br>Interviews conducted as part of a 1966 Canadian pilot project produced a nonrespondent vacancy rate 120 percent higher than the respondent vacancy rate. However, the nonrespondents were disproportionately very small firms; possible bias due to underrepresentation of small firms is dealt with separately below. (Dominion Bureau of Statistics, Part III.)<br>Calculations based on the very large amount of *JVS* microdata for the six survey occasions during February, March, and April, 1976, showed that correcting for nonresponse bias increased the estimated number of vacancies by an average of 8 percent. Under the assumption that respondents and nonrespondents are the same in average size, this implies a nonrespondent vacancy rate 20 percent higher than the respondent vacancy rate. In fact, large firms have somewhat higher response rates than small firms, so that the true difference between nonrespondent and respondents' vacancy rates is somewhat greater than 20 percent. |
| 3) Overrepresentation of large firms and underrepresentation of small firms in the reporting sample may cause an underestimate of the vacancy rate.<br>The *JOLTS* vacancy estimates were inflated by 15 percent to correct for this problem. All the other figures are based on stratified estimators. | According to unpublished tabulations supplied by Carol Utter of BLS, the U.S. manufacturing new hire rate estimated for March 1977 was 12 percent higher when recalculated to reflect stratification by 8 firm size groups. The bias in vacancy rates based on an unstratified estimator should be similar to the bias in new hire rates based on an unstratified estimator so long as the average duration of job openings is similar for small and large firms. |

TABLE 1—CONTINUED

| | |
|---|---|
| 4) Firms just being born will not be represented in the sampling frame. A priori, one might expect such firms to have very high vacancy rates.<br>All the available vacancy estimates were inflated by 5 percent to correct for this problem. | Between March of 1976 and 1977, universe employment grew faster than employment of firms in the labor turnover survey panel. Unpublished tabulations supplied by Carol Utter of BLS show that adjusting for firm births using the assumption that the omission of new firms from the panel caused this discrepancy raises the estimated March 1977 new hire rates by 4 percent. The downward bias in estimated new hire rates will equal the downward bias in comparably estimated vacancy rates if the average duration of job openings is the same at new and old firms.<br>Calculations based on Canadian new hire and employment data suggest that missing new firms might lead to an understatement of 2.5 percent in the estimated vacancy rate. (David Gower, 1970a.) |
| 5) Particularly in professional occupations, hiring may occur without any preceding active recruitment.<br>All the available vacancy estimates were inflated by 10 percent to correct for this problem. | Approximately 200 employers in Massachusetts and Utah kept vacancy diaries during the month of March 1980. Of 1134 persons hired by these employers, no more than 28 were walk-ins. (U.S. Department of Labor, 1981a, ch. 5.) |
| 6) There may be "discouraged vacancies" which employers would like to fill, but for which they have given up recruiting.<br>All the available vacancy estimates were inflated by 10 percent to correct for this problem. | As a part of the BLS 1979–80 pilot project, employers in Florida and Texas were asked to report both positions for which they were currently recruiting and positions which they had given up trying to fill. Adding in the discouraged vacancies increased the estimated vacancy rate by an average of 3 percent in Florida and by an average of 6 percent in Texas. (U.S. Department of Labor, 1981a, Appendix A.) |
| 7) Vacancy rates may be higher in the middle of the month, when the unemployment rate is measured, than at the end of the month.<br>To correct for this problem, the *JVS* vacancy estimates were inflated by 5 percent and the U.S. vacancy estimates were inflated by 10 percent. | During the first 9 months of *JVS* data collection (from the end of January through mid-October of 1970), the end-of-month and midmonth vacancy estimates exhibited a clear sawtooth pattern, with the end-of-month figures averaging roughly 9 percent lower than the midmonth figures. (Gower, 1970b.) |

program were fairly inclusive in their coverage, with the important exception that none of the surveys made any effort to cover the self-employed.[11] From the point of view of designing a vacancy survey, this exclusion seems quite reasonable, insofar as it is unclear that self-employment vacancy is even a meaningful concept. However, by assuming that the vacancy rate in the covered sectors equals the vacancy rate in the excluded sectors, one is implicitly imputing the existence of a substantial number of self-employment vacancies and thereby most likely exaggerating the true overall vacancy rate. For the U.S. *JOLTS* program, the relevant question is whether the manufacturing sector vacancy rate typically exceeds or falls below the vacancy rate in other sectors. In the other available data sources, the reported manufacturing sector vacancy rate tended to be higher than the rate for the nonmanufacturing sector, sometimes substantially so.[12] Assuming that the parts of the economy covered by and excluded from the U.S. *JOLTS* program have equal vacancy rates thus may well lead to overstating the overall vacancy rate for that time period.

---

[11] Data from the *Current Population Survey* (*CPS*) imply that the share of total employment accounted for by the self-employed and unpaid family workers averaged 9.6 percent between January 1970 and December 1980. By comparison, agricultural wage and salary workers accounted for an average of only 1.5 percent and private household workers for an average of only 1.7 percent of total employment. See U.S. Department of Labor (1981b, p. 156).

### C. *Unemployment Rate Data*

Before the adjusted vacancy rate estimates just described could be used for the purpose at hand, it was necessary to obtain unemployment rate estimates for the geographic areas and time periods covered by the various vacancy surveys. The only SMSA unemployment rates available to be matched with the 1964–66 BES/BLS pilot project vacancy data were monthly figures produced using the BLS "handbook" or "70 step" method. Overall national unemployment rates were used in conjunction with the national *JOLTS* data. The BLS supplied monthly unemployment rates based on the *Current Population Survey* (*CPS*) covering the period from 1970 to the present for the states of Minnesota, Wisconsin, Florida, Massachusetts, Texas, and Utah; these rates were matched with the Minnesota and Wisconsin *JOLTS* vacancy data and with the new BLS pilot project vacancy data. Finally, unemployment rate data based on the monthly *Canadian Labor Force Survey*, a household survey similar to the *CPS*, were obtained for the quarters covered by the Canadian *JVS*.

All of the available unemployment rate figures are based on the standard definition:

$$UR = [U/(U+E)] \times 100, \tag{3}$$

where $UR$ represents the overall unemployment rate, $U$ represents number unemployed, including persons without work during the entire survey reference week who 1) would like a job and have actively searched for one within the past four weeks, 2) are on temporary layoff, or 3) are due to start a new job within 30 days, and $E$ represents total employment in the economy. The unemployment numbers undoubtedly miss some people without a job who would like work, since they exclude anyone who has become convinced that no work is available and has for that reason given up active job search. During the period from the beginning of 1970 through the end of 1980, the number of discouraged workers (defined as persons who say they would like to work but have not searched within the past four weeks because they are convinced they will be unable to find a job) averaged almost 15 percent as large as the pool of persons officially counted as unemployed.[13] Furthermore, to have been counted as a vacancy, a job need only have been available on a single day, whereas to be counted as unemployed a person must have been out of work during an entire week. The unemployment rate would be at least somewhat higher if it were based on a one-day rather than a one-week reference period.

### D. *Vacancy Rates Associated with Different Unemployment Rates*

Table 2 reports the estimated corrected-for-bias vacancy rates associated with various levels of the unemployment rate. Two things seem noteworthy about the numbers which are presented. First, the estimated vacancy rates are surprisingly low. Except for the figures based on the BLS 1979–80 pilot project data, the vacancy rate estimates in the cells for unemployment rate levels above 5.0 percent all imply the existence of three or more unemployed people for every vacant job. The BLS 1979–80 pilot project data indicate that number of vacancies approximately equaled number unemployed for unemployment rates between 3.6 and 4.0

---

[12] Only the Wisconsin *JOLTS* program produced estimated vacancy rates for manufacturing which averaged below the relevant estimated overall vacancy rates. In the data produced by the BES/BLS 1964–66 pilots, the Minnesota *JOLTS*, the BLS 1979–80 pilots, and the Canadian *JVS*, the estimated manufacturing vacancy rates averaged above the estimated overall vacancy rates.

[13] See U.S. Department of Labor (1981b, pp. 119 and 142).

TABLE 2—ESTIMATES OF THE VACANCY RATE AT VARIOUS UNEMPLOYMENT RATE LEVELS

| Unemployment Rate Range | BES/BLS 1964–66 Pilot Projects | U.S. *JOLTS* Program | Minnesota *JOLTS* Program | Wisconsin *JOLTS* Program | BLS 1979–80 Pilot Projects | Canadian *JVS* Program |
|---|---|---|---|---|---|---|
| | Oct. 1964–Apr. 1966 | Apr. 1969–Dec. 1973 | Jan. 1972–Dec. 1979 | Jan. 1976–Dec. 1979 | Mar. 1979–June 1980 | Jan. 1971–Dec. 1978 |
| 1.6–2.0 | 3.3 ($N=3$) | – | – | – | – | – |
| 2.1–2.5 | 3.8 ($N=5$) | – | – | – | – | – |
| 2.6–3.0 | 2.6 ($N=9$) | 3.0 ($N=1$) | 3.0 ($N=2$) | – | – | – |
| 3.1–3.5 | 2.1 ($N=8$) | 2.5 ($N=5$) | 2.8 ($N=12$) | – | – | – |
| 3.6–4.0 | 1.9 ($N=3$) | 2.8 ($N=2$) | 2.4 ($N=17$) | 2.1($N=4$) | 3.8 ($N=4$) | – |
| 4.1–4.5 | 2.0 ($N=5$) | 1.9 ($N=8$) | 2.4 ($N=11$) | 1.8 ($N=12$) | 2.2 ($N=1$) | – |
| 4.6–5.0 | 2.2 ($N=4$) | 1.9 ($N=9$) | 2.1 ($N=17$) | 1.8 ($N=12$) | 4.2 ($N=4$) | 1.9 ($N=2$) |
| 5.1–5.5 | 1.5 ($N=1$) | 1.4 ($N=15$) | 1.7 ($N=16$) | 1.6 ($N=8$) | 3.5 ($N=5$) | 1.4 ($N=5$) |
| 5.6–6.0 | 1.6 ($N=2$) | 1.2 ($N=8$) | 1.8 ($N=8$) | 1.7 ($N=4$) | 4.3 ($N=3$) | 1.2 ($N=2$) |
| 6.1–6.5 | 1.2 ($N=2$) | 1.1 ($N=7$) | 1.4 ($N=5$) | 1.5 ($N=5$) | 3.5 ($N=4$) | 1.1 ($N=5$) |
| 6.6–7.0 | – | 0.8 ($N=2$) | 1.4 ($N=5$) | 1.7 ($N=1$) | 3.3 ($N=3$) | 0.9 ($N=7$) |
| 7.1–7.5 | – | – | 1.6 ($N=2$) | 1.3 ($N=1$) | – | 0.5 ($N=1$) |
| 7.6–8.0 | – | – | 1.3 ($N=1$) | 1.3 ($N=1$) | – | 0.8 ($N=7$) |
| 8.1–8.5 | – | – | – | – | – | – |
| 8.6–9.0 | – | – | – | – | – | 0.7 ($N=2$) |
| 9.1–9.5 | – | – | – | – | – | 0.5 ($N=1$) |

*Sources:* The available vacancy rate data and my corrections to the raw numbers are described in the text. The six sets of unadjusted vacancy rate figures were taken from the following places: Harrington (undated); *Employment and Earnings*; unpublished *JOLTS* tabulations for Minnesota and Wisconsin; U.S. Department of Labor (1981a); and Statistics Canada (1979). The matching unemployment rate data came from : Harrington (undated); unpublished *CPS* tabulations; and Statistics Canada (1980). None of the data used were seasonally adjusted.

percent; all of the other sets of vacancy rate estimates suggest that the number of available jobs has equaled or exceeded the number of unemployed persons only when the unemployment rate has fallen below 3 percent. Second, except in the column containing data for the BLS 1979–80 pilot projects, there appears to be a very strong negative relationship between the vacancy rate and the unemployment rate; that is, when the unemployment rate is high, the vacancy rate tends to be low, and vice versa. This inverse relationship has been labeled the Beveridge curve by other authors. The exact position of the Beveridge curve will depend upon the structure of labor market institutions, particularly those which operate to match workers with jobs, and on the degree to which workers' skills are compatible with job requirements.[14]

## E. *The Unemployment-to-Vacancy Ratio*

Given data on the vacancy rate and the unemployment rate, one can calculate the unemployment-to-vacancy ratio:

$$U/V=[UR/(100-UR)]\Big/[VR/(100-VR)], \tag{4}$$

where $U$ equals number unemployed, $V$ equals number of vacancies, $UR$ is the unemployment rate, and $VR$ is the vacancy rate. The unemployment rate and vacancy rate observations for each survey were used to calculate the mean unemployment-to-

---

[14] Bent Hansen (1970) develops a model of the labor market which implies that one should expect an inverse relationship between the vacancy rate and the unemployment rate. My dissertation (ch. 2) presents evidence that the U.S. Beveridge curve shifted outwards during the 1970's. There may also be regional differences in the position of the Beveridge curve. Because the relationship between the vacancy rate and the unemployment rate is not fixed, all inferences drawn in this paper regarding specific numbers are restricted to the time periods and areas covered by the relevant data.

TABLE 3—UNEMPLOYMENT VS. VACANCY AND JOB SEEKER VS. VACANCY ESTIMATES

| | BES/BLS 1964–66 Pilot Projects | U.S. *JOLTS* Program | Minnesota *JOLTS* Program | Wisconsin *JOLTS* Program | BLS 1979–80 Pilot Projects | Canadian *JVS* Program |
|---|---|---|---|---|---|---|
| | Oct. 1964– Apr. 1966 | Apr. 1969– Dec. 1973 | Jan. 1972– Dec. 1979 | Jan. 1976– Dec. 1979 | Mar. 1979– June 1980 | Jan. 1971– Dec. 1978 |
| Calculations Based Directly on Survey Data: | | | | | | |
| Mean unemployment rate | 3.6 | 5.0 | 4.8 | 5.0 | 5.3 | 6.7 |
| Mean vacancy rate | 2.4 | 1.6 | 2.1 | 1.7 | 3.7 | 1.0 |
| Mean unemployment-vacancy ratio | 1.9 | 3.9 | 2.7 | 3.2 | 1.6 | 8.4 |
| Mean job seeker-vacancy ratio | 1.5 | 3.3 | 2.2 | 2.6 | 1.3 | 6.8 |
| Calculations using U.S. Unemployment Rates for Period Spanned by Survey: | | | | | | |
| Mean unemployment rate: | 4.5 | 5.0 | 6.4 | 6.6 | 6.2 | 6.4 |
| Mean vacancy rate: | | | | | | |
| *VR* on *UR* model | 2.0 | 1.6 | 1.5 | 1.4 | 3.7 | 1.1 |
| ln *VR* on ln *UR* model | 1.8 | 1.6 | 1.5 | 1.5 | 3.5 | 1.1 |
| 1/*VR* on 1/*UR* model | 1.8 | 1.7 | 1.6 | 1.5 | 3.3 | 1.2 |
| Mean Unemployment-Vacancy Ratio: | | | | | | |
| *VR* on *UR* model | 2.4 | 3.8 | 5.9 | 5.0 | 1.7 | 7.5 |
| ln *VR* on ln *UR* model | 2.6 | 3.8 | 4.6 | 4.9 | 1.8 | 7.3 |
| 1/*VR* on 1/*UR* model | 2.6 | 3.9 | 4.4 | 4.8 | 1.9 | 7.4 |
| Mean Job Seeker-Vacancy Ratio: | | | | | | |
| *VR* on *UR* model | 2.0 | 3.1 | 4.8 | 4.3 | 1.4 | 6.1 |
| ln *VR* on ln *UR* model | 2.1 | 3.1 | 3.8 | 4.1 | 1.5 | 6.0 |
| 1/*VR* on 1/*UR* model | 2.2 | 3.2 | 3.6 | 4.1 | 1.6 | 6.1 |
| Implied Unemployment Rate at Which Vacancies Equal Unemployment: | | | | | | |
| *VR* on *UR* model | 2.8 | 2.8 | 2.8 | 2.2 | 3.8 | 2.2 |
| ln *VR* on ln *UR* model | 2.7 | 3.1 | 3.0 | 2.5 | 3.7 | 3.2 |
| 1/*VR* on 1/*UR* model | 2.5 | 3.3 | 3.1 | 2.8 | 3.7 | 4.3 |
| Implied Unemployment Rate at Which Vacancies Equal Job Seekers: | | | | | | |
| *VR* on *UR* model | 3.3 | 3.3 | 3.4 | 2.9 | 4.5 | 2.9 |
| ln *VR* on ln *UR* model | 3.2 | 3.4 | 3.4 | 3.1 | 4.4 | 3.5 |
| 1/*VR* on 1/*UR* model | 3.0 | 3.5 | 3.4 | 3.2 | 4.3 | 4.4 |

vacancy ratio for the relevant areas and time periods; these figures are reported in the top panel of Table 3.[15] Even though the ratios presented are based on unemployment rate numbers that exclude some persons who would like to work and vacancy rate numbers that very likely exaggerate the true overall availability of jobs, the estimated mean unemployment-to-vacancy ratios are consistently much greater than one.

Since the unemployment rates in the areas covered by the available vacancy survey data have tended to be below those for the United States as a whole, it also seemed of interest to ask what the unemployment-to-vacancy ratio would have been in these areas had the unemployment rate followed the same path as was observed for the entire United States. Given the apparent inverse relationship between the vacancy rate and the unemployment rate, one would expect a larger number of unemployed persons per vacant job at higher unemployment rates. The first step towards addressing this question was to summarize the vacancy rate/unemployment rate relationship for each of the six available sources of data. Because no particular functional form was an obvious choice on a priori grounds, three sets of Beveridge curve regressions were estimated: *VR* as a linear function of *UR*; ln *VR* as a linear function of ln *UR*; and 1/*VR* as a linear function of 1/*UR*. Except for those using the BLS 1979–80 pilot project data, all of the estimated models fit the data quite well.[16] Given an unemployment rate, the parameter estimates from any of these equations can be manipulated to yield an associated vacancy rate. The monthly U.S. unemployment rates for the period spanned by each survey were introduced into the Beveridge curve equations to get estimated vacancy rate values; the means of these unemployment rate and vacancy rate values are reported in the second panel of Table 3. Vacancy-to-unemployment ratios were then calculated using the actual monthly U.S. unemployment rates and the associated predicted vacancy rate values, and an average of these ratios was taken; these averages are also reported in the second panel of Table 3. The calculations imply that the vacancy-to-unemployment ratio would have averaged about 2.5 in the BES/BLS pilot project cities between October 1964 and April 1966 had those cities' unemployment rates moved with the U.S. unemployment rate, which averaged 4.5 percent over this period. The U.S. unemployment rate averaged 5.0 percent between April 1969 and December 1973; the U.S. *JOLTS* regressions imply that on average over this period there were almost 4.0 unemployed persons for every vacant job. The Minnesota *JOLTS* models, the Wisconsin *JOLTS* models, and the Canadian *JVS* models all imply that had these areas' unemployment rates paralleled the U.S. unemployment rate (which averaged roughly 6.5 percent over each of the relevant time periods), the average unemployment-to-vacancy ratio over the period covered by the relevant survey would have been roughly 5.0 or greater. While the BLS 1979–80 pilot project data suggest a much lower average unemployment-to-vacancy ratio than any of the other survey data, even that average is above 1.5.

---

[15] Table 3 contains both *unemployment* vs. vacancy comparisons and *job seeker* vs. vacancy comparisons; the job seeker figures are discussed in the next two subsections of the text.

[16] Excluding those based on the BLS 1979–80 pilot project data, the $R^2$s for the *VR* on *UR* models averaged .45; for the ln *VR* on ln *UR* models, they averaged .48; and for the 1/*VR* on 1/*UR* models, they averaged .44. The $R^2$s for the BLS 1979–80 pilot project models were

How low would the unemployment rate have had to fall for the number of vacancies to equal the number of unemployed persons? Excluding the BLS 1979–80 numbers, the average implied by the estimated Beveridge curve regressions is 2.9 percent. Only the regressions based on the BLS 1979–80 data consistently imply any higher unemployment rate and even those regressions suggest the unemployment rate would have to fall below 4 percent before the number of available jobs would equal the number of unemployed persons.[17]

### F. *Adjusting for the Inclusion of Persons not Seeking Jobs in the Unemployment Rate Data*

The conclusions reached thus far might be criticized on the grounds that the definition of unemployment is not symmetric with the definition of a job vacancy. In particular, jobs which are to be filled by recall from temporary layoff are not counted as vacancies, whereas persons who have been temporarily laid off are counted as unemployed. More fundamentally, it could be argued that we ought to be interested in the *job seeker*-to-vacancy ratio rather than in the *unemployment*-to-vacancy ratio.

The BLS tabulates number of job seekers for the United States as a whole, where the number of job seekers is defined as total number unemployed minus number on temporary layoff minus number due to start jobs within 30 days. These data are reported monthly; from January 1970 through December 1980, the number of job seekers thus defined averaged 83 percent as large as the number of unemployed persons.[18] The BLS job seeker figures very likely understate the actual number of persons who are looking for work since they exclude everyone on temporary layoff. Available evidence strongly indicates that many in this excluded group

---

all below .02. Lack of consistency in the data collection procedures employed in this pilot project effort may at least partially explain the poor fits.

[17] Any conclusions based on the BLS 1979–80 pilot project regressions are suspect since the unemployment rate variables in those models assumed unreasonably small and statistically insignificant coefficients.

[18] The BLS actually draws a distinction between what they call temporary layoff, which includes workers who expect to return to their employers within 30 days, and indefinite layoff, which includes everyone else on layoff who indicates a possibility of returning to their original employers. Following previous research, the term "temporary layoff" here refers to both of the above groups. The job seeker data for the period indicated were supplied by BLS.

do engage in active job search. Data from a supplement to the May 1976 *CPS* questionnaire indicate that persons on temporary layoff spent an average of 18.3 hours per month searching for a job; among the entire pool of unemployed, the average was 24.9 hours. The same data showed that persons on temporary layoff used an average of 2.5 search methods, compared to an overall average of 3.4 search methods. Moreover, data from the linked May, June, July, and August 1976 *CPS* files suggest that a very high proportion of those on temporary layoff do not return to their original employers.[19] Thus, it is my belief that the figures on the job seeker-to-vacancy ratio presented in Table 3 can be viewed as offering lower-bound estimates of the relevant true job seeker-to-vacancy ratios.[20]

### G. *The Job Seeker-to-Vacancy Ratio*

Job seeker data were not available for most of the areas and time periods covered by the available vacancy data. However, the relationship between the job seeker rate and the unemployment rate could be estimated fairly precisely using monthly U.S. data for the period from January 1970 through December 1980.[21] Job seeker rates were then created treating this estimated relationship as a prediction equation and plugging in the unemployment rates previously matched with the available vacancy data. These were used to calculate job seeker-to-vacancy ratios. The average values of these estimated job seeker-to-vacancy ratios appear in the top panel of Table 3. The estimated average job seeker-to-vacancy ratios are roughly 15 to 20 percent smaller than the comparable average unemployment-to-vacancy ratios.

What would the average job seeker-to-vacancy ratios in these areas have been had their unemployment and job seeker rates moved together with the overall U.S. rates? Monthly job seeker-to-vacancy ratios were calculated using the vacancy rates derived above (by inserting U.S. unemployment rates into the estimated Beveridge curve equations) and the actual values of the U.S. job seeker rate for the period spanned by each survey.[22] The middle panel of Table 3 reports the averages of these ratios. Again, the resulting ratios are roughly 15 to 20 percent smaller than the comparable unemployment-to-vacancy ratios. Excluding the BLS 1979–80 pilot project estimates, mean unemployment rates of 4.5 to 5.0 percent are associated with job seeker-to-vacancy ratios of 2.0 to 3.0 and mean unemployment rates near 6.5 percent are associated with job seeker-to-vacancy ratios above 4.0. Even the BLS 1979–80 pilot project numbers imply that at an unemployment rate near 6.0 percent, there would have been many more people actively seeking work than available jobs.

How low would the unemployment rate have had to fall for the number of vacancies to equal the number of job seekers? The Beveridge curve regressions described above specify the vacancy rate as a function of the unemployment rate; the relationship between the job seeker rate and the unemployment rate has also been estimated using data for the United States as a whole. Given that the vacancy rate equals $f(UR)$ and the job seeker rate equals $g(UR)$, where $UR$ is the unemployment rate, one can solve for the value of $UR$ such that the number of vacancies equals the number of job seekers. Excluding the BLS 1979–80 data, the average unemployment rate implied by these calculations is 3.3 percent. Even the calculations based on the BLS 1979–80 data imply that the unemployment rate would have to fall below 4.5 percent before the number of available jobs would equal the number of persons seeking work.[23,24]

---

[19] These figures are discussed in Kim Clark and Lawrence Summers (1979).

[20] One could conceivably go even further than is done here with removing people from the group officially counted as unemployed. Some of those included in the BLS job seeker count could be engaging in job search activities solely because they wish to maintain their unemployment insurance eligibility; some may be only marginally attached to the labor force. However, even if such additional adjustments were viewed as appropriate, there is no obvious way to objectively determine their magnitude.

[21] The job seeker rate was specified as a linear function of the unemployment rate and its square; the estimated regression had an $R^2$ of .94.

[22] Job seeker data are not available for months prior to January 1970. Values for these months were imputed based on the estimated relationship described in fn. 21.

## II. Alternative Estimates of the Unemployment-to-Vacancy Ratio

The preceding conclusions are based on vacancy rate data collected directly from surveyed establishments. An alternative route for obtaining unemployment-to-vacancy (or job seeker-to-vacancy) ratio estimates would be to start with data on new hire rates and on the average duration of a completed vacancy spell. In a steady-state situation:

$$NHR \cdot d_v = VR, \tag{5}$$

where *NHR* represents the new hire rate, $d_v$ represents the average length of time a job remains vacant, and *VR* represents the vacancy rate. In this section of the paper, data on job tenure from four January *CPS* files are used to estimate the new hire rate and available evidence on job vacancy durations is presented. The resulting new hire and duration information is combined to yield estimates of the vacancy rate for four points in time. These vacancy rates are similar in magnitude to the relevant rates obtained earlier. The results of this approach are thus supportive of the conclusion that the unemployment-to-vacancy and job seeker-to-vacancy ratios have typically been substantially above one.

### A. *Using Tenure Data to Estimate the New Hire Rate*

In January of 1963, 1968, 1973, and 1978, the *CPS* questionnaire contained a question regarding job tenure. Individuals were first asked whether they had been employed during the week containing the 12th day of the month. Those who said yes were asked when they had started the job at which they were employed. If hiring occurred at a uniform rate during the month of January, then it should be approximately true that

$$NHR = (31/k) \cdot (N_1/N_2), \tag{6}$$

where *NHR* is the new hire rate, $k$ is the date of the last day in the week including January 12, $N_1$ is the number of persons who said they started their jobs on a January date during or prior to the reference week, and $N_2$ is the total number of persons employed during the reference week. Using data from unpublished BLS tabulations, the relevant overall new hire rates were calculated to be as follows: January 1963, 6.5 percent; January 1968, 6.4 percent; January 1973, 4.7 percent; and January 1978, 6.2 percent.[25]

Are these numbers reasonable? One problem with the approximate formula described above is that newly hired persons who left their jobs prior to the survey reference week will not be counted. On the other hand, persons who were hired at an earlier date but did not start their new jobs until after January 1 *will* be counted. A priori, neither of these offsetting biases seems likely to be especially serious. One way to check the reasonableness of the new hire rates derived from the January *CPS* data would be to compare similarly computed new hire rates for manufacturing sector wage and salary workers to the manufacturing sector new hire rates obtained under the labor turnover program. However, some of the same downward biases that likely affected the *JOLTS* vacancy rates, in particular, the four problems of mail vs. interview responses differing, nonresponse bias, over-

---

[23] As noted in fn. 17, any conclusions based on the BLS 1979–80 pilot project regressions must be viewed with suspicion.

[24] An argument could be made that the vacancy vs. job seeker figures presented in Table 3 ought to have been calculated based on the underlying relationship between the vacancy rate and the job seeker rate, rather than based on the underlying relationship between the vacancy rate and the unemployment rate. However, as noted in the text, data on job seekers covering the same time periods and geographic areas as the available vacancy data could not be obtained. An alternative set of figures were prepared using Beveridge curve regressions which contained predicted job seeker rates as described in fn. 21 rather than the unemployment rate; the final results were virtually indistinguishable from those actually presented.

[25] Edward Sekscenski of BLS supplied the tabulations from which these numbers were derived.

representation of large firms, and nonrepresentation of firms just being born, would also be expected to affect the labor turnover program new hire rates. Using the correction factors developed earlier to deal with these four problems would suggest that the true new hire rate may be approximately 1.56 times as large as the new hire rate estimated based on the labor turnover panel reports. The *CPS*-based new hire rates for manufacturing sector wage and salary workers were thus compared to labor turnover program manufacturing sector new hire rates multiplied by a factor of approximately 1.56. The two resulting sets of numbers were as follows: January 1963, 3.7 percent (*CPS*-based) vs. 3.0 percent (labor turnover program); January 1968, 4.5 vs. 4.7 percent; January 1973, 3.6 vs. 5.5 percent; and January 1978, 4.5 vs. 3.9 percent.[26] The two sets of numbers appear to correspond fairly closely, at least with regard to their average level; the *CPS*-based set averages 4.1 and the inflated labor turnover program set averages 4.3.

### B. *Information on Job Vacancy Durations*

Two sources of information on job vacancy durations are available. The first is a Canadian study which relied upon questions asked during the interview phase of the *JVS*; the second is the vacancy diary project undertaken in the states of Massachusetts and Utah during March 1980.

The Canadian duration data were collected from *JVS* large employer interview forms (for units with 20 or more employees) for the period from the third quarter of 1971 through the first quarter of 1973. During the interview phase of the *JVS*, large employers were asked: 1) when did you last hire somebody?; and 2) when did you initiate action to fill the job for which this person was hired? Based on the answers to these questions, researchers at Statistics Canada estimated the mean length of time job vacancies remained open to be approximately 10 working days.[27] To the extent that large firms are more bureaucratic in their hiring process than smaller firms, this estimate may overstate the mean duration for all vacancies. Data on the duration of job openings was also collected as part of the BLS vacancy diary project carried out during March 1980 with the cooperation of Massachusetts and Utah employers. The employers were supposed to list the date they had begun recruiting for all jobs that were open at the beginning of the month, and then to record all recruiting and hiring activity which occurred during the month. The official report on the project contains a table showing percentage of jobs open 0 days, 1–7 days, 8–14 days, 15–30 days, 31–60 days, and 61 or more days, tabulated for all jobs filled or cancelled during the month.[28] Calculating mean job vacancy durations under the assumption that the midpoint of each interval represents the length of time it took to fill each job in the interval yields an estimated mean duration of 10.66 calendar days for Massachusetts and 8.07 calendar days for Utah. Calculations done by a state employment service staff member in Utah based on the actual duration of each completed job vacancy spell show an average duration of all filled or cancelled jobs of 6.76 calendar days, slightly under the 8.07 calendar days estimated using the interval data contained in the official report.[29]

Based on the available information, the average job vacancy remains open is somewhere between 5 and 15 calendar days. It is important to note that how long it takes to fill job openings may vary considerably over the business cycle. Both of the sources of information on duration discussed above apply to places and time periods where the unemployment rate was between 5 and 6 percent. One would expect shorter durations in situations with a higher unemployment rate and longer durations in situations with a lower unemployment rate.

---

[26] The relevant unadjusted labor turnover program new hire rates can be found in *Employment and Earnings*.

[27] See Winn Oughtred (1973).

[28] See U.S. Department of Labor (1981a, ch. 5).

[29] See Utah Department of Employment Security.

TABLE 4—ESTIMATED VACANCY FIGURES BASED ON TENURE DATA FROM THE JANUARY *CPS* FILES

| | Jan. 1963 | Jan. 1968 | Jan. 1973 | Jan. 1978 |
|---|---|---|---|---|
| Unemployment Rate | 6.6 | 4.0 | 5.5 | 7.0 |
| Implied New Hire Rate Based on Proportion of Employed Persons Who Started Job on a January Date During or Prior to the *CPS* Reference Week | 6.5 | 6.4 | 4.7 | 6.2 |
| Implied Vacancy Rate Assuming a Mean Vacancy Duration of: | | | | |
| 5 calendar days | 1.1 | 1.0 | 0.8 | 1.0 |
| 10 calendar days | 2.1 | 2.1 | 1.5 | 2.0 |
| 15 calendar days | 3.2 | 3.1 | 2.3 | 3.0 |
| Implied Unemployment-Vacancy Ratio Assuming a Mean Vacancy Duration of: | | | | |
| 5 calendar days | 6.6 | 4.0 | 7.6 | 7.5 |
| 10 calendar days | 3.3 | 2.0 | 3.8 | 3.7 |
| 15 calendar days | 2.2 | 1.3 | 2.5 | 2.5 |
| Implied Job Seeker-Vacancy Ratio Assuming a Mean Vacancy Duration of: | | | | |
| 5 calendar days | 5.6 | 3.2 | 6.5 | 6.2 |
| 10 calendar days | 2.7 | 1.6 | 3.2 | 3.1 |
| 15 calendar days | 1.8 | 1.1 | 2.1 | 2.0 |
| Best Guess about True Mean Vacancy Duration (calendar days) | 5 | 15 | 10 | 5 |
| Implied Vacancy Rate | 1.1 | 3.1 | 1.5 | 1.0 |
| Implied Unemployment-Vacancy Ratio | 6.6 | 1.3 | 3.8 | 7.5 |
| Implied Job Seeker-Vacancy Ratio | 5.6 | 1.1 | 3.2 | 6.2 |

### C. *Estimated Vacancy Figures*

Table 4 reports estimated vacancy rates, unemployment-to-vacancy ratios, and job seeker-to-vacancy ratios for January 1963, January 1968, January 1973, and January 1978. Three sets of figures are reported, the first assuming a mean vacancy duration of 5 calendar days, the second a mean duration of 10 calendar days, and the third a mean duration of 15 days. The 5-day average duration figure is probably the most appropriate for January 1963, with an unemployment rate of 6.9 percent, and for January 1978, with an unemployment rate of 7.0 percent. The best mean duration estimate for January 1973, when the unemployment rate was 5.5 percent, is probably 10 calendar days. The 15-day average duration figure is probably closest to accurate for January 1968, when the unemployment rate was 4.0 percent.

The calculations reported in Table 4 were intended to provide a check on the results obtained earlier. The figures arrived at seem to be quite consistent with the conclusions drawn from Tables 2 and 3. The most straightforward comparison to make between these new estimates and those presented earlier is to look at the vacancy rate associated with given values of the unemployment rate. Each of the vacancy rates in the row at the bottom of Table 4 is very similar to the vacancy rates from the appropriate positions in Table 2. Of the estimated unemployment-to-vacancy ratios in Table 4, only that associated with an unemployment rate of 4.0 percent is below 2.0. The unemployment-to-vacancy ratio associated with an unemployment rate of 5.5 percent is roughly 4.0, and those associated with unemployment rates near 7.0 percent are closer to 7.0. The corresponding job seeker-to-vacancy ratios average approximately 15 percent smaller.

While the calculations just described are admittedly rough-and-ready, it is reassuring that the results obtained based on two quite different approaches to estimating the vacancy rate, one via adjustment of the reported vacancy rates to correct for downward bias and the other via combining information on the new hire rate with information on the duration of job vacancies, should yield such consistent results.

## III. Conclusion

I have examined herein available data relevant for assessing how the number of persons wanting and/or seeking work has compared to the number of available jobs. Taken as a whole, the information compiled suggests that both the number of unemployed persons and the number of job seekers have consistently been much larger than the number of vacant jobs. If it could be assumed that the vacancy rate/unemployment rate relationships observed in the available survey data mirrored the vacancy rate/unemployment rate relationship prevailing in the United States over the same time period, reasonable estimates would be that there were roughly 2.5 unemployed persons for every vacant job during the middle 1960's, an average of close to 4.0 unemployed persons per vacant job during the early 1970's, and an average of 5.0 or more unemployed persons for every vacant job during the latter part of the 1970's. The number of persons counted as actively seeking work is somewhat smaller than the total number of unemployed persons, so that the comparable average job seeker-to-vacancy ratios would be 15 to 20 percent smaller.

What does my central finding that there are typically many more unemployed persons than job openings imply for macroeconomic theory and policy? It strongly suggests that measures such as training programs or increased job service funding designed to improve the process whereby unemployed workers are matched with available jobs, while perhaps a good idea for other reasons, cannot be expected to have any dramatic effect on the aggregate unemployment rate. If, for example, there are four unemployed persons for every vacant job, the aggregate unemployment rate would be only 25 percent lower, even if every available position could be filled *instantaneously* with an unemployed individual; the unemployment rate reduction which improvements in the matching process could *in fact* be expected to produce would be considerably smaller, since one cannot realistically hope to reduce the average duration of a vacancy to anywhere near zero. In most situations, large reductions in the aggregate unemployment rate will only be achieved if more jobs can be created. Lowering the minimum wage would lead to the creation of at least some new jobs, but available empirical evidence suggests this effect would be small.[30] Perhaps encouraging the use of shorter work weeks would be one appropriate policy response. The most standard approach to new job creation is obviously aggregate demand stimulation.[31]

Nothing said thus far should be taken as implying that having the same number of jobs vacant as there are persons unemployed is necessarily optimal. Suppose for the moment that policymakers could select freely any point along a given Beveridge curve. The optimal vacancy rate/unemployment rate combination will depend upon the marginal social costs associated with unemployment and with job vacancies. The social cost of another person becoming unemployed should equal the value of foregone production minus the value of time spent in job search minus the value of leisure time (possibly zero or even negative if there is physical and/or psychological stress associated with being out of work). The social cost of an added job vacancy should reflect at least the following: the cost to goods purchasers of their orders not being filled promptly; the cost to services purchasers who must wait in a longer queue; and the cost of any increase in inflationary pressure resulting from the existence of the vacancy.[32] One might be able to say, for example, that a given amount of aggregate demand stimulation could be expected to bring the number of job vacancies into rough equality with the number of unemployed persons; determining whether such stimulus would be desirable requires additional information. Assessing the relative social costs associated with job vacancies vs. unemployment would thus seem to be an interesting topic for future research.

---

[30] See Charles Brown, Curtis Gilroy, and Andrew Kohen (1982) for a good summary of available evidence.

[31] It is conceivable that improving the matching process and thereby lowering recruiting costs could induce employers to make more jobs available. These potential openings should be in large measure captured by the "discouraged vacancy" and "walk-in hire" adjustments made to the published vacancy data, as described earlier in this paper. In any event, this effect seems likely to be small.

[32] This framework for comparing the costs of unemployment and of vacancies is drawn from Robert J. Gordon (1973).

Finally, the evidence I have presented says nothing about whether and how best to facilitate those reductions in unemployment that conceivably could be achieved via pairing currently existing jobs with currently unemployed individuals. Are the vacant jobs unfilled because they are unattractive? Or are they unfilled because the unemployed lack the skills required to perform them? At what cost could the existing vacancy rate/unemployment rate tradeoff be improved? These, too, would be interesting questions for future investigation.

This paper has documented an important fact about the U.S. labor market: the number of persons seeking work seems typically to greatly exceed the number of available jobs. Additional work on the topic of the relationship between job vacancies and unemployment could further enhance our understanding of the nature of unemployment.

## REFERENCES

**Abraham, Katharine,** "Vacancies, Unemployment and Wage Growth," unpublished doctoral dissertation, Harvard University, 1982.

**Armknecht, Paul A.,** "Job Vacancies in Manufacturing, 1969–1973," *Monthly Labor Review*, August 1974, *97*, 27–33.

**Brown, Charles, Gilroy, Curtis and Kohen, Andrew,** "The Effect of the Minimum Wage on Employment and Unemployment," *Journal of Economic Literature*, June 1982, *20*, 487–528.

**Clark, Kim B. and Summers, Lawrence H.,** "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13–60.

**Friedman, Marvin,** "Comment," in *The Measurement and Interpretation of Job Vacancies*, New York: National Bureau of Economic Research, 1966, 132–37.

**Gordon, Robert J.,** "The Welfare Cost of Higher Unemployment," *Brookings Papers on Economic Activity*, 1:1973, 133–95.

**Gower, David,** (1970a) "Undercount of Vacancies Caused by New Establishments," mimeo., July 1970.

______, (1970b) "The Effect of the Reference Date on the Job Vacancy Estimates," mimeo., December 1970.

**Hansen, Bent,** "Excess Demand, Unemployment, Vacancies and Wages," *Quarterly Journal of Economics*, February 1970, *84*, 1–23.

**Harrington, Paul,** "Collection of Job Vacancy Statistics in Minnesota and Wisconsin," mimeo., April 1978.

______, "The Labor Department's Experience in the Collection and Analysis of Job Vacancy Statistics: The Experimental Programs of 1964–1966," mimeo., undated.

______ **and Milligan, William,** "A Brief Review of the Canadian Job Vacancy Survey," mimeo., undated.

**Oughtred, Winn,** "Job Vacancy Duration," mimeo., 1973.

**Dominion Bureau of Statistics, Department of Manpower and Immigration,** "Report of Committee on Job Vacancies Feasibility Survey," mimeo., December 1966.

**President's Committee to Appraise Employment and Unemployment Statistics,** *Measuring Employment and Unemployment*, Washington: USGPO, 1962.

**Statistics Canada,** *Canadian Job Vacancy Survey: Technical Appendix*, Catalogue 71–521 Occasional, Ottawa: Information Canada, 1972.

______, *Annual Report on Job Vacancies 1978*, Catalogue 71–203 Annual, Ottawa: Minister of Supply and Services Canada, 1979.

______, *Historical Labour Force Statistics 1979*, Catalogue 71–201, Ottawa: Minister of Supply and Services Canada, 1980.

**Subcommittee on Economic Statistics of the Joint Economic Committee,** 89th Congress, 2nd. Session, *Job Vacancy Statistics*, Washington: USGPO, 1966.

**The New York Times,** March 17, 1981, B10;

January 6, 1983, A24.

**The Wall Street Journal,** April 6, 1965, 16; July 10, 1969, 16; July 13, 1973, 6.

**U.S. Department of Labor,** *BLS Handbook of Methods,* Bulletin 1711, Washington: USGPO, 1971.

______, *Employment and Earnings,* Washington: USGPO, various issues.

______, (1981a) "Job Openings Pilot Program," mimeo., June 1981.

______, (1981b) *Employment and Training Report of the President,* Washington: USGPO, 1981.

**Utah Department of Employment Security,** "Utah Job Openings Survey Research Evaluations FY 1980," mimeo., December 1980.

# The Economies of Massed Reserves

*By* JAMES G. MULLIGAN*

The economy of massed reserves, first mentioned by E. A. G. Robinson (1958, pp. 26–27), is now firmly established in the industrial organization literature as an example of a potential plant-level scale economy. The massing of reserves results in a savings in the proportion of required reserves to expected output as the scale of a facility increases. Examples of such reserves are bank tellers, specialized tools and equipment, repairmen, inventories, spare parts, and checkout lines. The potential economies from massing reserves are dependent on the underlying stochastic processes governing the supply and/or demand of the service provided. Theoretical justification for the economies of massed reserves has usually been based on an appeal to the law of large numbers. For example, Donald Hay and Derek Morris wrote "the law of large numbers makes the number of breakdowns more predictable in a plant using a large number of machines, so that the number of stand-by maintenance staff need not be increased in proportion to size" (1979, p. 44).

This paper demonstrates that the law of large numbers does not provide a theoretical explanation of massed-reserves scale economies. Instead, it is shown that the expected economies of massed reserves can be calculated from the steady-state properties of well-known queuing models. Queuing models apply to any case where a servicing input is held in reserve to cope with the stochastic nature of the market or production process. In addition, a method for quantifying the expected economies of queuing processes is developed. The magnitude of these expected economies is shown to be significant and calculable. For example, the expected economies inherent in a widely applied queuing model are shown to exceed 23 percent in some cases.

Section I establishes the irrelevance of the law of large numbers as a theory of scale economies. Section II provides a formal method for determining the scale economies inherent in the most often mentioned massed-reserves example: machine repair. Section III shows that the same method can be used to calculate scale economies for the remaining massed-reserves examples: specialized repair tools and equipment; checkout counter clerks, bank tellers, and other service personnel; capacity; and inventories and spare parts. In addition, it is demonstrated that the same method can be used to calculate scale economies for some multiproduct operations.

## I. The Law of Large Numbers and Economies of Scale

The law of large numbers, stated briefly, refers to the statistical property that the sample mean approaches the true mean of a distribution as the number of randomly drawn observations approaches infinity. A formal definition, known as the Strong Law of Large Numbers, is as follows: If a common distribution of the independent, identically distributed random variables $X_1, X_2, \ldots$ has a finite first moment, $\mu$, then a sequence of averages, $Y_n = \Sigma X_n / n$, converges to $\mu$ with probability one.[1]

If the expected value of $X$ is $E(X)$, $nE(X)$ is the expected value of the sum, $S_n = \Sigma X_n$. The sum has a large variance, $n \operatorname{var} X$, that indicates a large dispersion as $n$ goes to infinity even when $\operatorname{var} X$ is finite. Assume that $X$ represents the machine breakdown rate of a production unit. Although the expected breakdown rate is $nE(X)$ for large $n$, the addition of an identical unit actually increases var$S_n$ by var$X$. If $E(X)$ is already known, there is no additional informational value attached to large size. Although proponents of the law of large numbers suggest that breakdowns become more predictable as $n$ increases, the large firm is likely to observe at any point in time values of $nX$ far from $nE(X)$. Since the variance around $nE(X)$, $n$var$X$, is the same for both a $n$-production-unit industry and a $n$-subunit firm, the large firm has no advantage. Since the law of large numbers does not guarantee that the large firm is anymore likely to reduce its expected number of breakdowns than a one-production-unit firm is, there is no obvious gain from reducing the proportion of standby repair personnel.

---

*Department of Economics, University of Delaware, Newark, DE 19711. I thank Kathleen Meier, Herbert Mohring, Marcel Neuts, John Schroeter, and the members of the University of Delaware Department of Economics Research Workshop for helpful comments and suggestions. Michael Caputo provided excellent research assistance. This paper was presented at the Winter 1982 meeting of the Econometric Society.

[1] See William Feller, pp. 259–60.

Although the law of large numbers has little to do with the traditional concept of scale economies, there are scale economies to be realized. These scale economies can be shown by considering the most commonly used massed-reserve example: machine repair. For expositional purposes, assume that there are $n$ identical firms each containing one worker and $k$ machines. The machines operate independently of workers until a machine breaks down. The worker is available to make repairs when necessary. Repair times will vary depending on the repair needed. Unless repairing a machine, the worker is assumed to be idle. The machines are assumed to be homogeneous in the sense that their breakdown and repair probability density functions are identical. Assume further that the parameters of the probability density functions are known by all firms, that the density functions do not change, and that they are not expected to change.

In the absence of diseconomies associated with transportation costs, span of management's control, and other related phenomena, a doubling of the scale of a firm is equivalent to combining two of the identical, independent firms. If these two firms, now formed into a combined firm of two workers and $2k$ machines, maintain their separate structure and each worker maintains separate care of his machines, there will be no economies from the merger. However, economies could result if otherwise idle workers are allowed to repair the machines of workers facing a queue of inoperative machines. The *potential* for economies will result only at a point in time when one worker has all of his $k$ machines working, and another worker has two or more machines needing repair. The economies are *realized* when an otherwise idle worker helps that worker by repairing a machine in the queue. They are thus due to the use of an otherwise idle resource. These joint events occur randomly. The magnitude of these economies will depend on the joint probability that an idle worker is matched with a worker with two or more machines in need of repair. Since this structural change reduces expected idle time, expected output per machine will increase for the larger firm. The probability of producing any given output can thus be maintained with fewer repairmen as the scale increases. Although there are fewer repairmen, each repairman spends less time idle. This scale economy has little to do with the law of large numbers or the ability of firms to predict the number of breakdowns more accurately. It is the result of a structural change that utilizes a resource (for example, repairmen) more fully.

## II. A Queuing Model of Scale Economies for Stochastic Markets

In this section the heuristic argument made in the previous section will be placed in the context of a well-known queuing model. While the literature on queuing theory is vast, surprisingly, it has almost completely ignored the inherent scale economies of queuing models. This neglect and the incorrect use of the law of large numbers suggest that a formal model of these economies is appropriate. This section uses the above machine-repair example to show the general application of queuing theory to modeling the economies inherent in stochastic markets and production processes, and develops a method for calculating these economies.

The machine-repair example will be used for three reasons: 1) machine repair is the most often mentioned example of massed reserves; 2) a queuing model used to characterize the stochastic nature of this process is both well known and relatively tractable;

and 3) the same queuing model can be applied to other nonmachine-repair examples. It is important to note, however, that the general approach shown here will apply to any situation where a serving unit is held in reserve to cope with the stochastic nature of the market or production process. Although the specific queuing model used will depend on the assumptions made about the frequency distributions of demand and/or supply, the method described is the same. In the next section, several nonmachine-repair examples will be provided to demonstrate the general usefulness of queuing models as both a theoretical explanation and a means for calculating scale economies.

The assumptions of the machine-repair example are now restated more formally: the production unit consists of a pool of $K$ potential arrivals to the system. In this example, $K$ is the number of machines. The machines operate independently of workers until a machine breaks down. There are $c$ servers (repairmen) available to monitor the system (machines) and serve (repair) any that arrive (break down). An individual's arrival rate (machine's breakdown rate) is assumed to be a Poisson random variable with mean $\lambda$. The service (repair) rate of a worker is also a Poisson random variable with mean $\mu$. The frequency of service time (repair length) is thus subject to a negative exponential distribution.[2]

For this model the steady-state probability of $n$ customers (machines) in need of service is given by

$$P_n = \begin{cases} \dfrac{K!}{(K-n)!n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (0 \leq n \leq c), \\[2ex] \dfrac{K!}{(K-n)!c^{n-c}c!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & (c \leq n \leq K), \end{cases}$$

where $P_0$ is found from the condition $\sum_{n=0}^{K} P_n = 1$. It has been shown in numerous sources that these probabilities yield the steady-state expected number of "down" machines, $D$, as

$$(1) \quad \sum_{n=0}^{c-1} n \frac{K!}{(K-n)!n!}\left(\frac{\lambda}{\mu}\right)^n P_0 + \sum_{n=c}^{K} n \frac{K!}{(K-n)!c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n P_0,$$

where

$$(2) \quad P_0 = \left[\sum_{n=0}^{c-1} \frac{K!}{(K-n)!n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K} \frac{K!}{(K-n)!c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}.$$

Setting $c$ equal to unity and $K$ equal to $k$ in (1) yields the steady-state number of down machines for the one-worker $k$-machine production unit as[3]

$$(3) \quad \sum_{n=1}^{k} \frac{nk!}{(k-n)!}\left(\frac{\lambda}{\mu}\right)^n \times \left[1 + \sum_{n=1}^{k} \frac{k!}{(k-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}.$$

Assume that steady-state expected output is proportional to the expected number of machines operating in a steady state. If there were no economies of scale inherent in this process and the scale of operation were increased with the server to potential arrival (repairman to machine) ratio held constant, expected output per machine would, by definition, be the same regardless of the size of the production unit. Output per machine for a production unit of $k$ machines and one worker would be identical to output per machine for a production unit of $ck$ machines and $c$ workers.

Yet, due to the nature of the Poisson breakdown process, a repairman will face a positive probability that no machines will require repair at a specific point in time. If the repairman is responsible for two or more machines, there will be a positive probability that two or more machines will need service at the same time. Combining two otherwise independent production units and allowing repairmen to assist one another will increase output per machine and worker, because of the matching of a repairman with no machines needing repair and a repairman with two or more machines needing repair. The economies from this matching process depend on $P(n=0)$, $P(n \geq 2)$, and the number of repairmen in the expanded unit. When $K/c$ and $\lambda/\mu$ decrease, $P(n=0)$ increases and $P(n \geq 2)$ decreases. As $K/c$ and $\lambda/\mu$ increase, $P(n=0)$ decreases and $P(n \geq 2)$ increases. Potential economies increase with the increase of potential interactions between repairmen with no machines requiring repair and repairmen with two or more machines needing repair. A formal analytic derivation of the number of machines and repairmen where these joint probabilities are maximized is a formidable task for this basic queuing model and is unnecessary. Numerical calculation of (1) for the full range of $K$, $c$, and $\lambda/\mu$ will yield exactly the same information and is a common approach used by queuing theorists when determining the properties of queuing models.

---

[2] This model is denoted in the queuing literature as $M/M/c/K$ or the finite source, $c$ server queue. A detailed discussion of queuing notation can be found in Donald Gross and Carl Harris, pp. 2–10.

[3] See Gross and Harris, pp. 118–19.

Expected steady-state output was calculated using equation (1) for the entire range of possible $\lambda/\mu$, $K$, and $c$. Since the economies will be maximized for each $K/c$ at a *unique* value of $\lambda/\mu$ and $c$, $K/c$ was chosen first and the economies calculated (in terms of a percentage increase in expected output per machine relative to the one-worker production unit holding $K/c$ constant) for the full range of $\lambda/\mu$ and $c$. The calculations verified that economies are maximized for each $K/c$ at a value exactly equal to $1+(\lambda/\mu)^{-1}$. Expressed equivalently, $(K-c)\lambda = c\mu$ or $\lambda/\mu = c/(K-c)$.[4] Movement of $K/c$ away from $1+(\lambda/\mu)^{-1}$ creates an excess of idle or busy workers due to the changes in $P(n=0)$ and $P(n \geq 2)$ and diminishes the importance of scale. For example, when $K/c$ is small relative to $1+(\lambda/\mu)^{-1}$, the probability that an idle worker's help is needed is small. Likewise, when $K/c$ is large relative to $1+(\lambda/\mu)^{-1}$, the probability that an idle worker is available is small.

Sample values of economies for various values of $K$, $c$, and $\lambda/\mu$ are presented in Table 1. As expected, Table 1 shows that for each $K/c$ ratio, deviations of $\lambda/\mu$ from $\lambda/\mu = c/(K-c)$ decrease the potential scale economies. As stated earlier, the reduction is due to an increase in the expected excess of either idle or busy workers as $\lambda/\mu$ deviates from $c/(K-c)$. This same phenomenon could be shown for deviations of $K/c$ from $1+(\lambda/\mu)^{-1}$. Although the firm cannot improve on expected output per machine at infinite scale by altering the $K/c$ ratio, uncertainty remains. The firm will still experience queues and idle time, but the firm has by such an organization coped with the uncertainty as much as possible.

Note that the profit-maximizing firm might not choose to operate at this ratio of $K/c$ because of relative factor prices. In the machine-repair context, $\lambda/\mu$ represents the technology (i.e., production function parameters) constraining the firms' decision. Thus the firm chooses the scale and the $K/c$ ratio subject to this technology and to factor prices. In the long run, the profit-maximizing firm may even have choices among competing technologies at various machine prices. In the short run, once the machines are in place, the firm can hire workers and thus still affect the $K/c$ ratio. The $K/c$ ratio chosen may not be the one predicted on scale economies alone, since relative factor prices dictate this decision. The scale arguments made here show that identical firms in a stochastic market can lower expected average cost at any given $K/c$ ratio by merging otherwise independent production units. For the machine-repair example, note in Table 1 that the full advantage of scale occurs at different scales depending on the $K/c$ ratio. The greatest potential occurs at $K/c$ equal to two and $\lambda/\mu$ equal to one. For $K$ equal to 1,000

---

[4] The importance of this result is tangential to the thesis of this paper. A paper giving a more detailed discussion of the result and a critical analysis of earlier, related work by David Levhari and Eytan Sheshinski, Moises Syrquin, and Kenneth Arrow, Levhari, and Sheshinski is available from the author upon request.

TABLE 1—SCALE ECONOMIES: PERCENT INCREASE IN OUTPUT PER MACHINE WHEN COMPARED TO THE ONE-WORKER REPAIR UNIT

A. $K/c=8$

| $\lambda/\mu$<br>Workers | .141 | .142 | .143[a] | .144 | .145 |
|---|---|---|---|---|---|
| 2 | 5.924 | 5.927 | 5.929 | 5.929 | 5.928 |
| 4 | 10.347 | 10.353 | 10.357 | 10.357 | 10.353 |
| 10 | 14.422 | 14.496 | 14.441 | 14.439 | 14.429 |
| 20 | 16.524 | 16.544 | 16.552 | 16.547 | 16.529 |
| 50 | 18.407 | 18.440 | 18.452 | 18.440 | 18.406 |

B. $K/c=16$

| $\lambda/\mu$<br>Workers | .065 | .066 | .067[a] | .068 | .069 |
|---|---|---|---|---|---|
| 2 | 4.666 | 4.682 | 4.689 | 4.686 | 4.675 |
| 4 | 8.094 | 8.129 | 8.142 | 8.132 | 8.098 |
| 14 | 12.041 | 12.127 | 12.150 | 12.108 | 11.997 |
| 35 | 13.703 | 13.844 | 13.878 | 13.790 | 13.574 |

C. $K/c=32$

| $\lambda/\mu$<br>Workers | .030 | .031 | .032[a] | .033 | .034 |
|---|---|---|---|---|---|
| 2 | 3.402 | 3.500 | 3.551 | 3.549 | 3.490 |
| 4 | 5.801 | 6.023 | 6.138 | 6.123 | 5.965 |
| 10 | 7.827 | 8.253 | 8.481 | 8.431 | 8.055 |
| 25 | 8.941 | 9.601 | 9.981 | 9.864 | 9.135 |

[a] The value of $\lambda/\mu$ where $\lambda/\mu$ equals $c/(K-c)$.

and $c$ equal to 500, the percentage increase is 23.7 percent. The scale needed to capture most of these economies is greatest for smaller $K/c$ ratios. For larger $K/c$ ratios, most of the scale economies are realized at relatively low values of $c$. Also note in Table 2 that as expected, the probability that any one worker is idle decreases when the scale increases.

The technique used for determining the scale economies for the machine-repair example can be summarized as follows: expected output for each value of $\lambda/\mu$, $K$, and $c$ is calculated using the steady-state properties of the appropriate queuing model. Maintaining the same $K/c$ ratio and value of $\lambda/\mu$, expected output for the one-server model is calculated. Comparisons of expected output per machine or equivalently per worker at different scale yields the potential economies. This technique can be used to determine the expected economies in any stochastic market where production units maintain a serving unit in reserve. The assumptions of Poisson arrivals and Poisson service can be changed to fit the stochastic process under consideration. Some examples are offered in Section III. Generalization of these assumptions, however, does increase the complexity of the steady-state properties of the queuing process.

The discussion to this point has assumed that reorganization is costless, a rather dubious assumption. In the machine-repair context, shirking is a possibility if workers are not given sole responsibility for specific machines.[5] A complete model of a market would incorporate the economies identified in this paper with well-documented diseconomies

---

[5] See Armen Alchian and Harold Demsetz.

TABLE 2—PROBABILITIES OF BEING IDLE AT DIFFERENT SCALES

A. $K/c = 8$

| $\lambda/\mu$ Workers | .141 | .142 | .143 |
|---|---|---|---|
| 2 | .13579 | .13272 | .12971 |
| 4 | .09971 | .09648 | .09333 |
| 10 | .06645 | .06306 | .05977 |
| 20 | .04931 | .04580 | .04243 |

B. $K/c = 16$

| $\lambda/\mu$ Workers | .065 | .066 | .067 |
|---|---|---|---|
| 2 | .11702 | .10940 | .10211 |
| 4 | .08810 | .08008 | .07249 |
| 10 | .06185 | .05331 | .04544 |
| 20 | .04854 | .03957 | .03155 |

such as transportation costs and management control.

### III. Additional Applications

This section considers the massing of other production inputs: (a) specialized repair tools and equipment; (b) checkout counter personnel, bank tellers, and other individuals providing service functions; (c) capacity; and (d) inventories and spare parts. While the specific application dictates the queuing model chosen, the general method is the same as shown in Section II. At the end of this section the possibility of scale economies for multiproduct operations is discussed.

#### A. *Specialized Repair Tools and Equipment*

A repair tool or a piece of equipment used to make repairs is conceptually the same as a repairman. Since the tool or piece of equipment is a server or repairman that remains idle until called upon to repair a machine, the queuing model used will be the same as that in Section II.

#### B. *Checkout Counter Personnel, Bank Tellers, and Others*

In this case, service is provided on demand by an individual facing a stream of customer arrivals. Instead of assuming that there are $k$ machines likely to break down, assume that there are $k$ identical, potential customers each likely to arrive with mean arrival rate $\lambda'$. As before, any assumption can be made concerning the nature of the arrival distribution. Assume for consistency of the exposition that this distribution is Poisson. Although there are $k$ potential customers, the steady-state expected number of customers in the system, that is, receiving service or waiting in line for service, will likely equal some smaller number $n$. Since a customer already in the system cannot arrive without first departing, the potential arrival rate is the arrival rate of a typical customer times the number of potential customers not already in the system, that is, $\lambda'(k-n)$. If the number of potential arrivals, $k$, is large relative to the expected number of customers in steady state, $n$, the expected arrival rate $\lambda'(k-n)$ is approximately equal to $\lambda' k$ without significant effect on the steady-state properties. This simplification is known as the infinite-source assumption and is commonly used for this type of application. The value $\lambda' k$ can be denoted $\lambda$, the mean arrival rate faced by the server. Service may be deterministic or stochastic. For example, assume that service is provided by $c$ servers according to a Poisson process with mean $\mu$. The appropriate model is known in the literature as $M/M/c$. When $c$ equals one, the steady-state number of customers in the system (i.e., in the queue or receiving service) is $\lambda/(\mu-\lambda)$.[6]

To determine the effect of scale changes, one would increase both the number of servers and the number of potential customers in the same proportion. A proportionate increase in the pool of potential customers will result in the same proportionate increase in $\lambda$, since $\lambda$ equals $\lambda' k$. Although changes in scale are normally defined as proportionate changes in all inputs, not proportionate changes in all inputs and customers, this assumption can be motivated as follows: consider two identical stores (for example, banks, supermarkets, etc.) located close enough geographically so that transportation costs of any kind are not an offsetting diseconomy. The arrival process is Poisson and the mean arrival rate is $\lambda$ for each store. Each store has one checkout counter. The two stores now merge and tear down the wall separating their stores. Assume that there is no defection of customers due to a personal distaste for either of the new partners. There is a well-known property that the sum of two independently distributed Poisson random variables of mean, $\lambda_1$ and $\lambda_2$, respectively, is a Poisson random variable with mean $\lambda_1 + \lambda_2$. (See Patrick Moran, p. 12.) Thus the resultant arrival process to the combined store will be Poisson with mean arrival rate $2\lambda$ and the number of servers will be two with each serving at mean rate $\mu$. The steady-state number of customers in the system for $c$ servers is

---

[6] See Gross and Harris, p. 53.

$$\frac{\lambda}{\mu}+\left[\frac{(\lambda/\mu)^{c}\lambda\mu}{(c-1)!(c\mu-\lambda)^{2}}\right]P_{0},$$

where[7] $$P_0=\left[\sum_{n=0}^{c-1}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^{n}+\frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^{c}\left(\frac{c\mu}{c\mu-\lambda}\right)^{-1}\right]^{-1}.$$

This merger presents the possibility that a customer waiting in one checkout line might be able to go to another counter serviced by an otherwise idle checkout clerk. This merger will reduce the expected time customers spend in the system and will reduce expected server idle time. Although numerous variations of this model exist depending on arrival and service distributions, priority of service and queue discipline, the method for the calculation of scale economies is the same. Interested readers are referred to Gross and Harris, among others, for the steady-state properties of these model variations.

### C. *Capacity Utilization*

F. M. Scherer et al. (1975) contend that sharing idle capacity is an example of massed reserves at the multiplant level. Since random fluctuations may result in excess demand at one plant when there is excess capacity at another, the sharing of capacity will lower the cost of meeting any given level of demand. The massing of capacity can be modeled with queuing theory as follows: the merging of two otherwise independent plants in the absence of transportation costs will have the same effect on the arrival rate as described in Section III, B. In this case $\lambda$ is the mean demand for the firm's product with each customer demanding one unit at a time.[8] Scale economies result in a reduction of customers' expected waiting time for a given capacity or a reduction in capacity needed to maintain a given level of customers' expected waiting time. In this case, the server is a unit of capacity, and the service it provides is a unit of output. A customer arrives and places an order. If capacity is available, the good is produced. Production may be deterministic or stochastic. Assume that production follows a Poisson process with mean rate $\mu$. The expected waiting time is $1/\mu$, once a unit of capacity has been allocated for the customer. Customers may have an additional, earlier wait in the queue before capacity is freed for their order. Increasing the plant's capacity amounts to an increase in the number of servers. Thus as long as capacity is available, orders will be processed in the same way that customers receive service at checkout counters. If demand and service are assumed Poisson, the appropriate model will be $M/M/c$ with scale changes defined as proportionate increases in $\lambda$ and $c$.

In addition to the variations mentioned earlier, one could consider the possibility that customers would balk and go elsewhere if the expected wait is too long. Arthur De Vany used a queuing model with customer balking to show that a natural monopolist will emerge in a market with fixed prices because of the reduction in customer waiting time as scale increases. Consumers were assumed to consider expected waiting time an indicator of quality. De Vany defined scale changes as proportionate increases in $\lambda$ and $\mu$ rather than $\lambda$ and $c$. De Vany and Thomas Saving state that there is little difference between the properties of a $M/M/c$ queuing model with $\mu$ equal to $\mu_0$ and the properties of a $M/M/1$ model with $\mu$ equal to $c\mu_0$. Their assumption does simplify the steady-state properties, but leads to an overestimate of the scale effects. A technical discussion of the conceptual and computational differences of these two models is available from the author upon request.

---

[7] See Gross and Harris, pp. 97–98.

[8] If customers demand more than one unit at a time, input to the system comes in the form of bulk or batch arrivals. See Gross and Harris, pp. 148–54, for a discussion of this assumption.

### D. *Inventories and Spare Parts*

Examples shown above (A; B; C) are direct applications of queuing theory. Although inventories of goods and spare parts often have been included in lists of massed-reserves examples, inventories do not provide a serving function. Inventories are consumed, once demanded, and are not available after use to meet subsequent demand. The inventory literature is extensive: In his survey article, Sumer Aggarwal wrote "the large number of research studies and models...do not in any way exhaust the possibilities of formulating the possible millions of additional models" (p. 476).[9] Although inventories do not provide serving functions, many situations can be modeled with the queuing theory. Other applications have been modeled with dynamic programming. Although a detailed treatment of all of these models is beyond the scope of this paper, the rationale for scale economies is shown to be the same as before.

Although C. F. Pratten, and John Haldi and David Whitcomb, among others, contend that the law of large numbers predicts a decrease in the ratio of required inventories to sales as scale increases, the law of large numbers is no more valid a justification for scale economies in this case than it was earlier. Firms hold inventories of goods and spare parts at positive storage cost in anticipation of stochastic demand. By sharing their inventories, two otherwise independent firms can maintain the same probability of having an item in stock while reducing expected storage costs. When one firm runs out of an item and a second firm has undemanded items in stock, the first can borrow from the second. This concept is the same as the sharing of capacity, repairmen, etc. The combined firm saves in two ways: it does not lose a customer that would otherwise have been turned away from the first subunit; and does not absorb the storage cost of maintaining the item in the other subunit's inventory.

While formal calculation of these economies will depend on the model used, queuing theory can be used for a number of applications. The notion of the server is different than before. Output will be the arrivals to the system. Unsold output (or unused spare parts) will be held in a queue. The queue is serviced by customers who remove the item at some rate. Customers can be handled simultaneously by assuming that there are places for $c$ customers at a time. Customers remove the good at mean rate $\mu$. If the customers arrive according to a Poisson process, and output or stocking of parts is a Poisson process, the appropriate model is again $M/M/c$. Here the probability that a server is idle is the probability that a customer arrives and does not find the good available. Scale economies can be calculated by increasing $\lambda$ and $c$ in fixed proportion and comparing expected idle time. If firms are concerned that they will lose customers when the good is not available, they will attach an expected cost to $P(n=0)$ and weigh it against the expected cost of holding positive inventory levels in steady state.

### E. *Economies of Scope*

Most firms produce more than one product. The possibility of scale economies for multiproduct firms, called economies of scope, has received considerable attention since the initial contributions of John Panzar and Robert Willig (1975). (See Elizabeth Bailey and Ann Friedlaender, 1982, for a recent survey.) Economies of scope are said to result when a multiproduct firm can produce a given amount of two different goods at a lower cost than the combined cost of two firms each specializing in one of the two goods. More formally stated, economies of scope result when $C(X_0, Y_0) < C(X_0, 0) + C(0, Y_0)$, where $C(X_0, Y_0)$ is the cost of the multiproduct firms production of $X_0$ and $Y_0$ amounts of goods $X$ and $Y$, respectively, and $C(X_0, 0)$ and $C(0, Y_0)$ are the costs of the two one-product firms. A potential source of these economies is the sharing or joint utilization of certain inputs.

---

[9] For additional surveys of the inventory literature, see Steven Nahmias and Edward Silver.

If the demand for these goods is stochastic and the input performs a serving function, the firm will experience the same expected economies as mentioned earlier. For example, assume that prices for two goods are fixed for some finite period of time and that demand is a Poisson process. Assume that demand for $X$ and $Y$ are independent with means, $\lambda_X$ and $\lambda_Y$, respectively. As shown earlier the demand faced by the firm will be a Poisson process with mean $\lambda_X + \lambda_Y$, if $X$ and $Y$ are homogeneous goods. If the firm uses the same process or equipment to produce $X$ and $Y$, the products may be perfect substitutes in production even though less than perfect substitutes to consumers. Assume that $X$ is produced by a firm with capacity $c$, and $Y$ is produced by another firm with capacity $c$. Assume that each unit of capacity provides output according to a Poisson process with mean $\mu$. Combining the two production processes will make the combined firm a $2c$ server system with each server providing service at rate $\mu$. Capacity is being shared by both products. Although these same economies could be captured by a similarly sized firm producing just $X$ or $Y$, diseconomies such as transportation costs may preclude the expansion in the production of the one good. Since doubling $\lambda_X$ may not be possible without substantial advertising or transport cost, the firm may find it more profitable to produce both $X$ and $Y$ and sell locally.

## IV. Summary

This paper has considered the potential scale economies for firms operating in a market governed by stochastic demand and/or supply, and challenges the commonly expressed view that these economies are a natural consequence of the law of large numbers. Instead, the steady-state properties of well-known queuing models are used to provide formal theoretical justification and a means of explicitly calculating the potential economies due to either the expansion of an existing firm or the merger of two or more firms. This research shows that potentially significant economies do exist. These economies are dependent on parameters of the queuing model specified. Economies result from the greater use of resources, such as repairmen, clerks, capacity, and inventories that would have otherwise spent a larger percent of time idle. These results suggest that in the absence of offsetting diseconomies larger firms would have a potential cost advantage in markets governed by stochastic demand and/or supply.

## REFERENCES

**Aggarwal, Sumer,** "A Review of Current Inventory Theory and Its Applications," *International Journal of Production Research*, July 1974, *12*, 443–82.

**Alchian, Armen and Demsetz, Harold,** "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

**Arrow, Kenneth, Levhari, David and Sheshinski, Eytan,** "A Production Function for the Repairman Problem," *Review of Economic Studies*, July 1972, *39*, 241–49.

**Bailey, Elizabeth and Friedlaender, Ann,** "Market Structure and Multiproduct Industries," *Journal of Economic Literature*, September 1982, *20*, 1024–48.

**De Vany, Arthur,** "Uncertainty, Waiting Time and Capacity Utilization: A Stochastic Theory of Product Quality," *Journal of Political Economy*, June 1976, *84*, 523–42.

______ **and Saving, Thomas,** "Product Quality, Uncertainty, and Regulation: The Trucking Industry," *American Economic Review*, September 1977, *67*, 583–94.

**Feller, William,** *An Introduction to Probability Theory and Its Applications*, Vol. I, New York: John Wiley & Sons, 1968.

**Gross, Donald and Harris, Carl,** *Fundamentals of Queuing Theory*, New York: John Wiley & Sons, 1974.

**Haldi, John and Whitcomb, David,** "Economies of Scale in Industrial Plants," *Journal of Political Economy*, August 1967, *75*, 384.

**Hay, Donald and Morris, Derek,** *Industrial Economics: Theory and Evidence*, Oxford: Oxford University Press, 1979.

**Levhari, David and Sheshinski, Eytan,** "A Microeconomic Production Function," *Econometrica*, May 1970, *38*, 559–73.

**Moran, Patrick,** *The Theory of Storage*, New York: John Wiley & Sons, 1959.

**Mulligan, James G.,** "Stochastic Production Functions and Queuing Theory," mimeo., University of Delaware, 1983.

**Nahmias, Steven,** "Perishable Inventory Theory: A Review," *Operations Research*, July-August 1982, *30*, 680–708.

**Panzar, John and Willig, Robert,** "Economies of Scale and Economies of Scope in Multi-Output Production," Economic Discussion Paper No. 33, Bell Laboratories, 1975.

**Pratten, C. F.,** *Economies of Scale in Manufacturing Industry*, Cambridge: Cambridge University Press, 1971.

**Robinson, E. A. G.,** *The Structure of Competitive Industry*, Chicago: University of Chicago Press, 1958, 26–27.

**Scherer, F. M.,** *Industrial Market Structure and Economic Performance*, 2d ed., Chicago: Rand McNally & Co., 1980, 83.

______ **et al.,** *The Economics of Multi-Plant Operations: An International Comparisons Study*, Boston: Harvard Economic Studies, 1975, 274–84.

**Silver, Edward,** "Operations Research in Inventory Management: A Review and Critique," *Operations Research*, July-August 1981, *29*, 629–43.

**Syrquin, Moises,** "Returns to Scale and Substitutability in the Repairmen Problem," *Econometrica*, September 1972, *40*, 937–41.

# American Economic Growth and the Voyage of Columbus

By R. PRESTON MCAFEE*

"Life, as we know it, does not exist."
*Swami Mhakhaphi*

Since the imaginative, pathbreaking, inventive analysis of Robert Fogel (1962), the counterfactual analysis has intrigued and scintillated a generation of economists. Fogel considered the state of the American economy in 1890, had the railroads never been invented. He found that less than 10 percent of the American output could be attributed to the single innovation of railroads, thus demonstrating irrevocably that the loss of trains would not derail the American economic juggernaut. However, no stationary equilibria exist. Indeed, Fogel's analysis showed that the American economy was resilient in ways not previously considered. Furthermore, Fogel's article had greater impact. After all, if the railroads did not exist in 1890, they probably do not exist now, for there is no record of them being invented in the last ninety-two years.[1] Nonexistent railroads are unsafe to use (L. M. Mijhak, 1835).

Since Fogel's analysis, the counterfactual has been successfully employed in a variety of other areas. A. Gibbons and O. Ho (1966) considered what would have transpired had the transatlantic railroad not been constructed at the turn of the century. They found that, with the necessarily higher costs of transportation, lighter than air vehicles would have been developed. Their conclusion was disputed by S. S. Vicente (1967) who proved conclusively that lighter than air vehicles are not possible, for no construction materials having sufficient rigidity are lighter than air. Nevertheless, X. O. Warbly (1968) analyzed the state of the world if "aeroplanes" were possible, and found they would not have been used anyway. Errors in Warbly's procedure were found by Muon (1970), who reevaluated the counterfactual using the Multiple Augmented Data Construction Assistance Package. Muon arrived at the same conclusion as Warbly, although the difference in cost and value of aeroplanes is reduced. In a clever twist, Shatter and Filth (1975) consider "what if Fogel had never written his article," and projected that economic historians would have turned to cocaine use instead of counterfactuals. Finally, Dromedary (1976) hypothesized that if Fogel were conceived on a train, he would have utilized more traditional methods in economic history.

The problem all of these articles share is that, by the time rail transportation is introduced, the U.S. economy had already produced the entrepreneurial class that paved the way for capital accumulation and growth. As a popular slogan of the 1890's asserted, "you can't keep a good robber baron down." Fogel demonstrates that railroads were not fundamental to the American economy, but does not consider deeper issues. In retrospect, it is obvious that Fogel was on the wrong track, for without railroads, entrepreneurs would have found other ways to bilk the government out of millions of square acres[2] of land. Thus, in demonstrating the invincibility of the American economic juggernaut, Fogel created a straw man.[3]

In order to perform a valid test of the invincibility of the American economic cornucopia, the counterfactual must predate the development of the celebrated entrepreneur and the waves of immigrants whose sweat was an important input into the production process. Consequently, it is hypothesized that, rather than stumble upon the two American continents, Columbus fell off the edge of the earth. Certainly this is a valid test, for if America were to be virtually unchanged, despite not being discovered, certainly the "American century" was inevitable. I choose the year 2000 as a target date, and compare America as it will be in the year 2000 to the way it would have been then, had Columbus fallen off the edge.

---

*Department of Economics, Social Science Centre, University of Western Ontario, London, Canada N6A 5C2. I thank Frank Bacon, Danny Bernoulli, Chuck Darwin, Pete Fermat, Chuck Fourier, Eddy Gibbon, Dave Hilbert, Joe Lagrange, Fred Nietzsche, and Bert Russell for valuable discussions. All errors are attributed to stultification resulting from Canada's climate and its first-best beers.

[1] There is one undocumented report of the steam locomotive being invented in the USSR in 1927, by Igor Antonvich. No report of rails being employed there surfaces, however, until 1949.

[2] There is some evidence that bilking the government out of round acres is sufficiently difficult that bilking would be reduced to $\pi/4$ of levels otherwise attained. This assumes no overlapping properties, however, and scale economies in honest graft.

[3] This is generally regarded as a despised type, except among Iowa children.

The paper proceeds in three parts. First, I construct a model of the earth, in which implications of Columbus falling off the edge can be tested. A new analytic procedure, the highly touted Fractured Reconstructive Autoerotic Projection Package with Econometrization (FRAPPE), from my 1983 study is used to detail the properties of the counterfactual world. These results are summarized in the third section. The fourth section offers conclusions and some proposals for further research.

Before formulating the model, a few possible objections may be laid to rest. Dribble (1980) has argued that America was not discovered by Columbus, but by another Italian with the same name. In a later paper, Dribble (1981) alleges that Isabella did not finance Columbus,[4] but that Ferdinand, dressed like the Queen of Spain, did. Readers should note that Dribble now drives a milk truck in Cleveland.

A more substantive issue was raised by Occam, who states (reprinted without permission):

> If the world were flat, Columbus would certainly not have observed masts rising over the horizon as ships came in. As a result, he would have not believed the earth to be round, and would never have sailed. Hence he could not have fallen off the edge. [1983, p. 17]

Fikus (1980) is less ambiguous on this point, stating that, in the presence of rational expectations, Columbus would have sailed only if he would reach India.[5] Milcfed (1981) counters that Columbus should sail if and only if he will discover something worthwhile, and thus it is sufficient for Columbus to sail if he feels it is worthwhile to find out what it feels like to fall off the edge. Occam (1985) will agree to this point, ecstatic over the elegance of the approach.

This shows that it is not contradictory to presume Columbus sailed away and fell off the earth. Certain topographical facts are implied by this statement, which form the core of the model developed in the next section.

## I. The Model

Let $\mathscr{X}$ be a subset of an uncountable cross product of the complex space, with the subset product topology generated from the lexicographic topology on complex space. Let $B$ represent the Borel sets on $\mathscr{X}$, generated from this topology, and $\mu$ a nonnegative measure on $B$. If $\mathscr{X}$ is compact, let $\mathscr{X}* = \mathscr{X}$, otherwise let $\mathscr{X}*$ be the one point compactification of $\mathscr{X}$ with induced metric $\mu*$. Extend $\mu*$ to the nonstandard reals, and denote this extended metric $*\mu*$, over $*\mathscr{X}*$, the nonstandard extension of $\mathscr{X}*$.

Choose a point $s \in *\mathscr{X}*$.[6] A *path* from $s$ to $x$ is a continuously differentiable mapping $p:[0,1] \rightarrow *\mathscr{X}*$ such that $p(0) = s$ and $p(1) = x$. Such a path is a *catastrophe in* $\mathscr{X}$ if $(\exists t)(0 < t < 1$ and $p(t) \in *\mathscr{X}* \setminus \mathscr{X})$. A catastrophe is a path leaving $\mathscr{X}$. Thus, if $\mathscr{X}$ is the earth's surface, a catastrophe is the event of leaving the earth's surface. In the context of this model, we may state the counterfactual hypothesis as follows:

$$(1) \qquad (\exists s \in \mathscr{X})(\exists x \in *\mathscr{X}*)$$

($p$ is a path from $s$ to $x \Rightarrow p$ has a catastrophe). It follows that (1) implies $*\mathscr{X}*$ is disconnected. Let $\mathscr{Y}$ be the largest connected subset of $\mathcal{X}$ containing $s$ in the discrete topology, and $*\mathcal{Y}*$ the one point compactification of the nonstandard extension of $\mathcal{Y}$.

---

[4] Which Columbus is meant is not clear.

[5] In point of fact, Columbus did sail, did not reach India, and the earth is not round anyway but ellipsoidal. Fikus correctly notes that the facts are irrelevant to a counterfactual argument.

[6] Chosen as it represents Spain.

Define an individual's state by his position in $*\mathcal{Y}*$ (location), his consumption rate $z \in \mathcal{Z}$, and a vector of attributes $v \in \mathcal{V}$.[7] Thus an individual is an element of $*\mathcal{Y}* \times \mathcal{Z} \times \mathcal{V}$. The state of the world is given by a vector of individual types, that is, an element in $\hat{W} \equiv \times_{\alpha \in \Gamma}(*\mathcal{V}* \times \mathcal{Z} \times \mathcal{V})$, where $\Gamma$ indexes all things.

The evolution of the earth can be given by a function $\Psi : \mathfrak{R} \times \hat{W} \rightarrow [0,1]$, where $\Psi(t, W)$ is the probability of $W$ at time $t$. There is some uncertainty in the function determining $W$ at the subatomic level, and thus statements concerning the actual world state have some variance.

Define $U$ to be the projection of $W$ onto the United States. Let $W_0$ be the actual state of the world in the year 2000, and $U_0$ the projection of $W_0$ on $U$. Let $\Phi(W) = \Psi(2000, W)$, given (1),[8] and define

$$W_1 = \int_{\hat{W}} W\Phi(W)\, d*\mu*.$$

That is, $W_1$ is the expected world state given the counterfactual (1). Then we have, if $U_1$ is the projection of $W_1$ onto $U$:

THEOREM: *In the coarse topology pseudometric,*[9] $\rho$, $\rho(U_0, U_1) = 0$.

PROOF:

For proof, send $8.99 to Proof, Box 666, London, Canada. (Canadians add $6.)

This shows that, despite Columbus falling off the edge, the United States is on average substantially unchanged, relative to the actuality. To get a grip on the real implications of the counterfactual, the most likely world state was generated, given the counterfactual, using FRAPPE, an artificial intelligence procedure that simulates the development of the world, given an initial state. Some of the results we found for this hypothetical world are listed in the next section. A full explanation of the FRAPPE technique can be found in my 1983 study.

## II. Results of the FRAPPE

The FRAPPE is a maximum-likelihood forecast procedure that takes as an input the initial state of the world and the stochastic differential equations governing state transformation. From this, FRAPPE constructs the most likely subsequent state. This is used as an input, and the process repeated. To find the state of the world on January 1, 1490, FRAPPE is used in reverse, calculating the most likely predecessor to the current state. Thus, to calculate the state of the world $W_0$ in 1490, everything about the world is inputted as a vector $W_2$ with dimension corresponding to the cardinality of the reals. Anti-FRAPPE is operated until the most likely initial state in 1490 is found. At this point, the initial state is altered so that the counterfactual hypothesis holds. FRAPPE is then used to find the world state on January 1, 2000. Simulations reveal that FRAPPE conclusions are insensitive to subatomic stochastic fluctuations for periods up to 500 years.[10] Thus, FRAPPE is valid for the objective of this paper, but not useful for testing the Big Bang Theory.

---

[7] There need not be finitely many attributes, although this saves on computational costs.

[8] The necessity of deriving the nonstandard complex measure space $*\mu*$ is obvious.

[9] A pseudometric on a space $S$ is a function $p : S \times S \rightarrow \mathfrak{R}$ satisfying:

(a) $(\forall x \in S)(\forall y \in S)(p(x,y) = p(y,x) \geqslant 0)$;

(b) $(\forall x \in S)(\forall y \in S)(\forall z \in S)(p(x,z) \leqslant p(x,y) + p(y,z))$.

A psuedometric maps leather into those cheesy shoes that look like vinyl.

[10] In particular, FRAPPE is useful in forecasting human events, weather, and the stock market for the next several hundred years. For the last reason, FRAPPE's mechanics must remain a trade secret. However, the details will be published in my 2010 study posthumously. Unfortunately, the common knowledge of FRAPPE would destroy FRAPPE's analytical ability, for the same reason undecidability crops up in the theory of algorithms, and thus FRAPPE will not be exploitable when generally known.

To summarize the FRAPPE results, I present a rough chronology of the events following Columbus' demise. While America is virtually unchanged in 2000 A.D., the rest of the world is drastically different.

After Columbus disappeared, rumors ran rampant that a Nirvana lay across the sea. Hundreds of ships were launched from Spain and Portugal to this Paradise, and with each ship, the promise of returning exacted. None returned, and Spain's power as a nation diminished with this burden on the treasury, and the Moors recaptured Spain in 1511.

Meanwhile, the industrial revolution gathered strength in the British Isles and Northern Europe. Fed by trade with India and China across the land, Italy, Palestine, and the Moslem countries prospered. An Englishman, F. Drake, experimented with a hot air balloon developed by a friend; the first channel crossing occurred in 1604. Without the Americas as an avenue of exploration, hardy types substituted balloon travel. Frustrated at the number of middlemen between Europe and China, France and England captured large portions of the Middle East in the "Great Balloon War" of 1712. This was followed by the "Second Balloon War" of 1731, in which the first motorized dirigible was used. The leader of the British forces, Margaret Attila "the Hen," was defeated at Qum in 1735, and the Europeans sued for peace.

Frustrated in eastward expansion, and intrigued by the studies of A. Monrovia concerning plate tectonics and "floating continents," the English turned their eyes toward newly discovered Australia. In a daring exploit, a team of Welsh divers unfastened Australia from the Pacific rim and, in dead of night, sneaked it past Indian and African shipping, to fasten it near the edge of the earth roughly 1,000 miles off Brittany.[11]

The new "continent of America," named for an aborigine phrase meaning "what?," was first used as a penal colony to house former British Lords and French royalty. These individuals eventually revolted, and the rest is history.

The rest of the world is perturbed in the FRAPPE analysis. For instance, individuals of the Indian subcontinent and native Americans were never given the same name, and certain confusions avoided. Moreover, the former did not acquire the statistical preeminence they are noted for, instead becoming the world's foremost proctologists. Japanese technical prowess was diverted away from electronics, automobiles, and processed seaweed toward military exports and grain, the latter grown hydroponically in 75-story farms.

Certain differences in the United States cropped up as well. California, placed next to the edge of the world, still had its earthquakes and flakes. However, Baltimore became an exciting city and a "den of iniquity," instead of the dreary middleclass drudge it is today.

### III. Conclusions and Research Proposals

This paper demonstrates that, whether or not Columbus discovered America, the United States would have been relatively unchanged. Thus, America is invincible to tinkering with history to a degree heretofore unsuspected.

It would be interesting to know if this is true in general. That is, is it the case that, for all $\varepsilon > 0$, a $\delta > 0$ exists so that, if history is perturbed by less than $\delta$, then America is perturbed by less than $\varepsilon$? In this case, America would be continuous, and hence integrable. One might then wonder if the United States converges in probability to what it will be, as time diverges.

Several related topics emerged during this analysis. Suppose Ponce de Leon (1513) had discovered the Fountain of Youth in Florida. Would the South have won the Civil War? What would have happened to the sales of grits and carpetbags in Boston?

If the earth had more than one moon, it is likely the space race would have turned out differently. Indeed, a general solution to the $m$ nation $n$ moon problem would shed light on certain general equilibrium welfare issues raised by Errow (1981). Errow demonstrated that if preferences are homothetic and moons are triangular, one should add them up.

---

[11] Travelling around the Cape, Australia was accidently inverted. Thus, the shape of the United States was not significantly altered.

Other industrial questions concern us. What would have transpired if one size didn't fit all? If your mileage didn't vary depending on how you drive? And if contents did not settle during shippage, would cardboard be elastic?

Two interesting counterfactual analyses are given by Deviant (1970) and Vildeside (1975). Deviant considers the effects on scholastic aptitude and economic growth if torture had become an accepted way of teaching undergraduates. Vildeside considers, in a partial equilibrium framework, necessary alterations in search theory if the searcher's physical attributes are endogenous, in the "putty-clay" sense.

Critics of counterfactuals have charged[12] that, given the inability to forecast the U.S. economy even months in advance, and the dearth of all but the most coarse data on historical events, it is absurd to expect to reconstruct how the economy would have been if it were different. There are three avenues of response to this, two theoretical and one practical. First, the inability to forecast in the modern economy is irrelevant, for the economy was simpler in the old days. Second, if the data are coarse, one can use the coarse topology[13] to analyze them. Thus, no theoretical impediments exist to applying counterfactuals, although it is interesting to suppose they did.

Finally, if there are countably many events, there are continuum-many subsets of events that can be negated and studied via the counterfactual argument. This produces a wealth of avenues of inquiry not seen in any other discipline.[14] Moreover, counterfactuals fall in the grand tradition of "assuming it away."

## REFERENCES

**Deviant, B. U.**, *Lessons of the Vivisectionists*, April 1970, not available in stores.

**Dribble, D. D.**, "$\nu$ Facts Concerning Columbus," June 1980, unpublished.

______, "Odd Practices of Spanish Kings, 1490–92," *Milk Deliverers Local No. 266 Important Notices*, April 1981, posted in pool hall.

**Dromedary, T. H.**, "Conceptions of Counterfactuals," *Journal of Irreparable Results*, March 1976, *22*, 17–68.

**Errow, M. B.**, "Partitioning the Moon: Optimal Geometric Structures," *Geopolitical Astronomy: A Journal of Theory and Practice*, February 1981, *7*, 1–32.

**Fikus, A. X.** "Was Columbus Rational?," *Journal for the Application of Algebraic Topology in Economic History*, April 1980, *17*, 4+2i–18+7i.

**Fogel, Robert**, "A Quantitative Approach to the Study of Railroads in American Economic Growth," *Journal of Economic History*, June 1962, *22*, 163–97.

**Gibbons, A. and Ho, O. O.**, *The Transatlantic Railroad*, Cleveland: Dufey Press, 1966.

**Leon, P. D.**, "Exotic Foundations of Florida," in A. E. Gator, ed., *Traveler's Guide to Florida*, St. Augustine: Coquina Press, 1513, 2017–137.

**McAfee, R. P.**, *Tales of Mountains and Shepherding*, French Lick: Zen Press, 1983.

______, *The Years 1980–2480, A Summary*, McAfee Press, 2010.

**Mijhak, L. M.**, "Railroads Don't Work and are Unsafe as Well," *National Enlightener*, June 5, 1835, 1–2.

**Milcfed, V. L.**, "Simple Paths to Joy," *Women's Christian Bowler Weekly*, June 5, 1981, *102*, 7–8.

**Muon, Ψ. Φ.**, "Warbly's 'Jets' Suck Air," in P. T. Scrotum, ed., *Metrizable Spaces in Physical Education: The Theory of Healthy Minds and Bodies*, Newark: North-Holland Tunnel Press, 1970, 10–41.

**Occam, R.**, "Are Rational Expectations Elegant? Some Revealing Examples," *The Econoclast*, April 1983, *186*, 17–18.

______, *Memoirs of a Two Edged Sword*, Byzantium: Elegance Press, 1985.

**Onan, M. E.**, "Positive Externalities in New

---

[12] Several such charges were made in a made-for-TV movie that aired last year. I cannot remember the film's name but it concerned teenage pregnancy.

[13] If $S$ is a space, the coarse topology is the set $\{\emptyset, S\}$. This topology is generated by the pseudometric that everywhere vanishes.

[14] For instance, there are merely countably many possible mathematical theorems in a countable language such as English.

Zealand Sheepherding and the Shortage of Gum Boots," *Royal White Paper Series on the Paucity of Women in the New Zealand Sheepherding Industry*, Christchurch, New Zealand, 1980.

**Shatter, J. O. and Filth, E. S.**, "Counter Fogel Counterfactual," *Recreational Pharmacology Quarterly*, May 1975, *21*, 1–10.

**Vicente, S. S.**, "An Intuitive Proof of the Nonexistence of the So-Called 'Aeroplane'," *Journal of Mathematical Epistemology*, August 1967, *72*, 82–100.

**Vildeside, W. W.**, "Stochastic Dominance and Endogenous Traits," *Annals of Deviant Behavior*, May 1975, *15*, 69–666.

**Warbly, X. O.**, "The Aeroplane in Economic and Military Applications: A Counterfactual Analysis," *Taiwan Journal of Economic Theory and Available Rental Housing*, September 1968, *12*, 100–21.

**Yucca, N. H.**, "The Cross Price Elasticity for Gum Boots and Lambs," Sydney Polytechnical Institute Paper Series, No. 465, 1967.

**Zeagram, J. E. and Sons**, "The Theory of the Best Fifth," *The Proceedings of the Crown Royal Conference on Game Theoretic Analysis of Sophisticated Procedures Converging to Being Blotto*, Montreal: Rye Press, 1982, 5–70.

# Uncertain Innovation and the Persistence of Monopoly

*By* JENNIFER F. REINGANUM*

A topic of long-standing interest in industrial organization is the effect of monopoly power upon incentives to engage in innovative activity. More recently, the concept of monopoly has been replaced with incumbency, so that explicit account may be taken of the existence of potential (and actual) rivals. According to F. M. Scherer:

> There is abundant evidence from case studies to support the view that actual and potential new entrants play a crucial role in stimulating technical progress, both as direct sources of innovation and as spurs to existing industry members. ...new entrants contribute a disproportionately high share of all really revolutionary new industrial products and processes. [1980, pp. 437–38]

In a recent paper published in this *Review* (1982), Richard J. Gilbert and David M. G. Newbery show that, because an incumbent firm enjoys greater marginal incentives to engage in *R&D* (under their assumption of deterministic invention), the incumbent firm will engage in preemptive patenting. Thus the industry will tend to remain monopolized, and by the same firm.

In this paper, I present a model in which an incumbent firm and a challenger engage in a game of innovation in which the inventive process is stochastic. I show that when the first successful innovator captures a sufficiently high share of the post-innovation market (i.e., when the innovation is sufficiently revolutionary), then in a Nash equilibrium the incumbent firm invests less on a given project than the challenger. Under an alternative specification, in a Nash equilibrium an incumbent firm conducts fewer parallel projects than would a challenger. In either case, the incumbent is less likely to patent the innovation than is the challenger.

The intuition for this result is relatively straightforward, at least in the case where the first successful innovator captures the entire post-innovation market. When the inventive process is stochastic, the incumbent firm continues to receive flow profits during the time preceding innovation. This period is of random length but is stochastically shorter the greater the firms' investments in *R&D*. Since a successful incumbent merely "replaces himself" (albeit with a more profitable product), the incumbent firm has a lower marginal incentive to invest in *R&D* than does the challenger.

Thus this paper provides a theoretical model that embodies Scherer's empirical observations: entrants stimulate progress both through their own innovative behavior and through their provocation of incumbent firms. Moreover, in equilibrium they contribute a disproportionate share of important innovations. I also attempt to isolate the causes of the discrepancy between my results and those of Gilbert and Newbery, and to integrate the two into a coherent theory consistent with the empirical observations.

## I. Related Literature

For simplicity, consider a case of cost-reducing innovation in an industry with constant returns to scale. Let $\bar{c}$ denote the incumbent firm's initial unit cost, and let $c<\bar{c}$ be the unit cost associated with the new technology. Let the relevant profit rates be $R=$ the current revenue flow to the incumbent firm; $\Pi(c)=$ the present value of monopoly profits using the new technology; also the present value of profits to the current incumbent if the incumbent receives a patent on the new technology; $\pi_I(c)=$ the present value of Nash-Cournot profits to the current incumbent firm if the challenger re-

---

*Associate Professor of Economics, Division of Humanities and Social Science, California Institute of Technology, Pasadena, CA 91125. I thank Ed Green, Roger Noll, John Roberts, and Louis Wilde for helpful comments. The financial support of the National Science Foundation is gratefully acknowledged.

ceives a patent on the new technology; and $\pi_C(c)$ = the present value of Nash-Cournot profits to the challenger if the challenger receives a patent on the new technology.

ASSUMPTION 1: *The functions* $\Pi(c)$, $\pi_I(c)$, *and* $\pi_C(c)$ *are continuous, and piecewise continuously differentiable. Moreover,* $\Pi(c)$ *and* $\pi_C(c)$ *are nonincreasing in c while* $\pi_I(c)$ *is nondecreasing in c.*

That is, if the incumbent patents the new technology, its profits will be lower the higher is the unit cost associated with the new technology. On the other hand, if the challenger patents the new technology (and the incumbent continues to use the old one), then the challenger's profits will be lower and the incumbent's higher the higher is the unit cost associated with the new technology.

*Definition* 1: The innovation will be termed *drastic* if $c \leqslant c_0$, where $c_0$ is assumed to exist and to be uniquely defined as the maximum value of $c$ such that $\pi_I(c) = 0$.

The important feature of the constant returns to scale assumption is that if profits are zero, so is output. Thus if $c \leqslant c_0$, then the current incumbent's output will be zero after the challenger patents the innovation. In this event, the challenger is a monopolist and $\pi_C(c) = \Pi(c)$. Note that $\Pi(c) \geqslant \pi_I(c) + \pi_C(c)$ with strict inequality whenever the innovation is not drastic.

*Remark* 1: Since $c < \bar{c}$, the present value of monopoly profits after innovation $\Pi(c)$ always exceeds the present value of monopoly profits without the innovation, $R/r$, where $r$ is the discount rate. Furthermore, $R/r > \pi_I(c)$ for all $c < \bar{c}$. This is because $R/r = \Pi(\bar{c}) > \pi_I(\bar{c}) \geqslant \pi_I(c)$ for all $c < \bar{c}$ by Assumption 1.

The following example illustrates the preceding discussion and Assumption 1. If the demand curve is linear, $P = a - bQ$, then the functions above are $\Pi(c) = (a-c)^2/4rb$, $\pi_I(c) = (a - 2\bar{c} + c)^2/9rb$ and $\pi_C(c) = (a - 2c + \bar{c})^2/9rb$, whenever the expressions in parentheses are nonnegative; otherwise the relevant value for the function is zero. Each of these functions is continuously differentiable except at the point at which the expression in parentheses becomes zero, and continuity is preserved at that point. The innovation is drastic whenever $c \leqslant c_0$, where $a - 2\bar{c} + c_0 = 0$. From this equality, it is easy to see that $\Pi(c_0) = \pi_C(c_0)$ and $\pi_I(c_0) = 0$.

Gilbert and Newbery argue as follows. If the inventive process is deterministic, then whoever is willing to bid most for the new technology receives the patent first with probability 1. The challenger will be willing to bid up to $\pi_C(c)$, while the incumbent will be willing to bid up to $\Pi(c) - \pi_I(c)$. Since $\Pi(c) \geqslant \pi_I(c) + \pi_C(c)$, with strict inequality for $c > c_0$, the incumbent preemptively patents the new technology. Only if the innovation is drastic will the incumbent and the challenger invest an equal amount. Consequently, preemption is the Nash equilibrium outcome in the bidding game. Thus the industry will remain monopolized and in the hands of the current incumbent.

This is clearly true when there is no uncertainty in the innovation process. The natural extension of this result to the case of uncertain innovation is that the incumbent is *more likely* to patent the innovation than is the challenger.[1] In the next section, a simple

---

[1] Gilbert subsequently argues; "Uncertainty in the invention process does not greatly change the deterministic analysis of preemption, provided $R$ and $D$ expenditures are sensitive to the expected returns and the established firm is no more averse to risk than rivals" (1981, p. 299). In a laudatory comment on the Gilbert paper, Richard Craswell continues: "Assuming any form of direct relationship between the amount spent on $R$ and $D$ and the likelihood of making the invention first, the incumbent will end up with the patent more often than not, and his monopoly will be maintained. In fact, the incumbent will usually end up with the patent even if he is less efficient at $R$ and $D$ than are his rivals, so long as his inefficiency does not completely negate the advantage due to his larger expenditure on $R$ and $D$" (p. 272).

To summarize, Gilbert and Craswell evidently believe that the result that the incumbent invests more than the challenger extends straightforwardly to the case of uncertainty. A somewhat different argument is offered by Gilbert and Newbery in support of the same basic claim: "Similar results are obtained when the assumption of a deterministic patent date is replaced by a more general stochastic function which describes the probabilmodel is presented which incorporates uncertainty. It is found that, for drastic innovations, the incumbent always invests less than the challenger, so that the incumbency changes hands *more often than not*. Due to the continuity of the equilibrium investment rates in the unit cost associated with the new technology, there will be an open neighborhood of $c_0$, representing innovations which are not drastic, for which the incumbent still invests less than the challenger. In Section III, I briefly report results from a more general model in which firms are allowed to choose both the number of parallel projects to undertake, and the rate of investment on each project. Again there is a nontrivial set of innovations which the incumbent is less likely to patent than is the challenger.

## II. A Model Incorporating Uncertainty

The model developed in this section is a generalization of that of Tom Lee and Louis Wilde (which is itself based upon a model by Glenn Loury). An incumbent and a challenger are simultaneously attempting to perfect a particular cost-reducing technology. Technological uncertainty takes the form of a stochastic relationship between the rate of investment and the eventual date of successful completion of the new technology. If $x_I$ represents the rate of investment of the incumbent, and $\tau_I(x_I)$ the random success date of the incumbent, then $Pr\{\tau_I(x_I) \leqslant t\} = 1 - e^{-h(x_I)t}$, for $t \in [0,\infty)$. Similarly, if $x_C$ and $\tau_C(x_C)$ represent the investment rate and the random success date for the challenger, then $Pr\{\tau_C(x_C) \leqslant t\} = 1 - e^{-h(x_C)t}$. The expected success date for firm $i$ $(i = I, C)$ is $1/h(x_i)$, where the function $h(.)$ is the hazard function used in much of the recent literature on patent races. In particular, following Loury, Lee and Wilde, and Partha Dasgupta and Joseph Stiglitz, I assume:

ASSUMPTION 2: *The hazard function* $h(.)$ *is twice continuously differentiable, with* $h'(x) > 0$ *and* $h''(x) < 0$ *for all* $x \in [0,\infty)$. *Furthermore,*

$$h(0) = 0 = \lim_{x \to \infty} h'(x).$$

*Thus the technology exhibits decreasing returns to scale.*

Suppose that the new technology is patentable so that the race ends with the first success. The expected profit to the incumbent for any pair of investment rates $(x_I, x_C)$ is

$$\begin{aligned} V^I(x_I,x_C) &= \int_0^\infty e^{-rt}e^{-(h(x_I)+h(x_C))t} \\ &\times \left[h(x_I)\Pi(c) + h(x_C)\pi_I(c) + R - x_I\right] dt \\ &= \left[h(x_I)\Pi(c) + h(x_C)\pi_I(c) + R - x_I\right] \\ &\Big/ \left[r + h(x_I) + h(x_C)\right]. \end{aligned}$$

That is, the incumbent receives $\Pi(c)$ at $t$ if the challenger has not yet succeeded and the incumbent succeeds at $t$; this event has probability density $h(x_I)e^{-(h(x_I)+h(x_C))t}$. The incumbent receives $\pi_I(c)$ at $t$ if the incumbent has not yet succeeded and the challenger succeeds at $t$; this event has probability density $h(x_C)e^{-(h(x_I)+h(x_C))t}$. Finally, the incumbent receives flow profits of $R$ and pays flow costs of $x_I$ so long as no firm has succeeded; this event has probability $e^{-(h(x_I)+h(x_C))t}$.

The challenger's payoff is analogous.

$$\begin{aligned} V^C(x_I,x_C) &= \int_0^\infty e^{-rt}e^{-(h(x_I)+h(x_C))t} \\ &\times \left[h(x_C)\pi_C(c) - x_C\right] dt \\ &= \left[h(x_C)\pi_C(c) - x_C\right] \\ &\Big/ \left[r + h(x_I) + h(x_C)\right]. \end{aligned}$$

The differences between these payoffs arise from the incumbent's current profit flow and

---

ity of invention at date $T$ conditional on a particular $R$ and $D$ plan. ...The argument is the same as before, except that the monopolist has to set up the correct number of rival research teams..." (p. 521). Thus the implied claim is that, if allowed to select the number of parallel projects to be undertaken, an incumbent firm would choose a larger number than would a challenger.

the fact that it shares the market in the event of successful innovation by the challenger.

*Definition* 2: A *strategy* for the incumbent (challenger) is an investment rate $x_I(x_C)$. The *payoff* to the incumbent (challenger) is $V^I(x_I, x_C)(V^C(x_I, x_C))$.

*Definition* 3: A *best response function for the incumbent* is a function $\phi_I:[0,\infty)\rightarrow[0,\infty)$ such that, for each $x_C$, $V^I(\phi_I(x_C), x_C)\geqslant V^I(x_I, x_C)$ for all $x_I\in[0,\infty)$. Similarly, a *best response function for the challenger* is a function $\phi_C:[0,\infty)\rightarrow[0,\infty)$ such that, for each $x_I$, $V^C(x_I, \phi_C(x_I))\geqslant V^C(x_I, x_C)$ for all $x_C\in[0,\infty)$. The best response functions will also depend upon the parameters $(c, R)$.

*Definition* 4: A strategy pair $(x_I^*, x_C^*)$ is a *Nash equilibrium* if $x_I^*=\phi_I(x_C^*)$ and $x_C^*=\phi_C(x_I^*)$. That is, each firm's investment rate is a best response to the other's.

The proof of the following proposition can be found in the Appendix.

PROPOSITION 1: *If* $h'(0)\geqslant\max(1/[\Pi(c)-R/r], 1/\pi_C(c))$, *then there exists a best response function*[2] *for the incumbent* $\phi_I(x_C; c, R)$ *which satisfies the first-order condition* $\partial V^I(\phi_I, x_C)/\partial x_I=0$ *and the second-order condition* $\partial^2 V^I(\phi_I, x_C)/\partial x_I^2<0$. *The function* $\phi_I$ *is continuously differentiable in its argument* $x_C$ *and continuous in the parameters* $c, R$. *Similarly, there exists a best response function for the challenger* $\phi_C(x_I; c)$ *which satisfies the analogous first- and second-order conditions, and is continuously differentiable in its argument* $x_I$ *and continuous in the parameter* $c$. *Moreover, there exists a pair of Nash equilibrium strategies* $x_I^*(c, R)$ *and* $x_C^*(c, R)$; *each is continuous in the parameters* $c, R$.

The first-order conditions which implicitly define the best response functions are

$$(1)\quad \partial V^I(\phi_I, x_C)/\partial x_I \propto [r+h(\phi_I)+h(x_C)][h'(\phi_I)\Pi(c)-1] - [h(\phi_I)\Pi(c)+h(x_C)\pi_I(c)+R-\phi_I]h'(\phi_I)=0;$$

$$(2)\quad \partial V^C(x_I, \phi_C)/\partial x_C \propto [r+h(x_I)+h(\phi_C)][h'(\phi_C)\pi_C(c)-1] - [h(\phi_C)\pi_C(c)-\phi_C]h'(\phi_C)=0.$$

Rearranging terms and noting the definitions of $V^I(\phi_I, x_C)$ and $V^C(x_I, \phi_C)$ yields

$$(3)\quad V^I(\phi_I, x_C)=[h'(\phi_I)\Pi(c)-1]\big/h'(\phi_I);$$

$$(4)\quad V^C(x_I, \phi_C)=[h'(\phi_C)\pi_C(c)-1]\big/h'(\phi_C).$$

*Remark* 2: Since the individual firm payoffs must be nonnegative, particularly when the firms play best responses, it follows that $h'(\phi_I)\Pi(c)-1\geqslant0$ and $h'(\phi_C)\pi_C(c)-1\geqslant0$.

LEMMA 1: $\partial\phi_I(x_C; c, R)/\partial x_C>0$ *and* $\partial\phi_C(x_I; c)/\partial x_I\geqslant0$. *Thus the existence of the challenger provokes the incumbent to invest more than it otherwise would on the innovation.*

PROOF:

By the implicit function theorem,

$$\partial\phi_I/\partial x_C=-[\partial^2 V^I(\phi_I, x_C)/\partial x_C\partial x_I]\big/[\partial^2 V^I(\phi_I, x_C)/\partial x_I^2].$$

The denominator is negative by the second-order condition. The numerator is

$$-h'(x_C)[h'(\phi_I)(\Pi(c)-\pi_I(c))-1] = -h'(x_C)h'(\phi_I)[V^I(\phi_I, x_C)-\pi_I(c)]$$

---

[2] The hypothesis of this proposition ensures that each firm's payoff function is initially nondecreasing in that firm's investment level. Without this assumption, it is possible that the firm may have two (widely separated) best responses.

by equation (3). Since $\phi_I$ is a best response to $x_C$, $V^I(\phi_I, x_C) \geqslant V^I(0, x_C)$; but $V^I(0, x_C) - \pi_I(c) = [R - r\pi_I(c)]/[r + h(x_C)] > 0$ by Remark 1. Thus $V^I(\phi_I, x_C) - \pi_I(c) > 0$ for all $x_C$. It follows that $\partial\phi_I/\partial x_C > 0$ for all $x_C$.

By the implicit function theorem,

$$\partial\phi_C/\partial x_I = -\left[\partial^2 V^C(x_I, \phi_C)/\partial x_I \partial x_C\right] \Big/ \left[\partial^2 V^C(x_I, \phi_C)/\partial x_C^2\right].$$

The denominator is negative by the second-order condition, while the numerator is $-h'(x_I)[h'(\phi_C)\pi_C(c) - 1]$, which is nonpositive by Remark 2.

LEMMA 2: *If the innovation is drastic and* $R > 0$, *then* $\phi_I(x; c, R) < \phi_C(x; c)$ *for all* $x, c$.

PROOF:

Recall that if the innovation is drastic, $\pi_C(c) = \Pi(c)$ and $\pi_I(c) = 0$. Then the only difference between equations (1) and (2), which implicitly define the best response functions $\phi_I$ and $\phi_C$, is the term $R$, representing current profit flows to the incumbent firm. If $R = 0$, and the innovation is drastic, then $\phi_I(x; c, 0) = \phi_C(x; c)$ for all $x, c$. Again using the implicit function theorem, we see that

$$\partial\phi_I/\partial R = -\left[\partial^2 V^I(\phi_I, x)/\partial R\, \partial x_I\right] \Big/ \left[\partial^2 V^I(\phi_I, x)/\partial x_I^2\right].$$

Since the denominator is negative and the numerator is $h'(\phi_I)$ which is positive, we have $\phi_I(x; c, R) < \phi_I(x; c, 0) = \phi_C(x; c)$ for all $R > 0$, and all $x, c$.

PROPOSITION 2: *If the innovation is drastic and* $R > 0$, *then in a Nash equilibrium, the incumbent invests less than the challenger; that is,* $x_I^*(c, R) < x_C^*(c, R)$.

PROOF:

Suppose, contrary to the proposition, that $x_I^*(c, R) \geqslant x_C^*(c, R)$. Then Lemmas 1 and 2 and the definition of a Nash equilibrium imply that

$$\begin{aligned} x_C^*(c, R) &= \phi_C(x_I^*(c, R); c) \\ &\geqslant \phi_C(x_C^*(c, R); c) \\ &> \phi_I^*(x_C^*(c, R); c, R) = x_I^*(c, R). \end{aligned}$$

But this is a contradiction. Thus $x_I^*(c, R) < x_C^*(c, R)$.

COROLLARY 1: *If* $R > 0$, *then there exists an open neighborhood of* $c_0$ *(which may depend on* $R$*), denoted* $N(c_0; R)$, *such that if the technology is not drastic, but* $c \in N(c_0; R)$, *then* $x_I^*(c, R) < x_C^*(c, R)$.

PROOF:

This follows immediately from Proposition 2 and the continuity of the Nash equilibrium investment rates $x_I^*(c, R)$ and $x_C^*(c, R)$ in the parameter $c$.

Thus I have concluded that for sufficiently radical innovations (i.e., for technologies in the set $N(c_0; R)$), it is precisely the assumption of certainty vs. uncertainty which is responsible for the discrepancy between my results and those of Gilbert and Newbery. To see the economics of the issue, consider what happens in my model with drastic innovation if the incumbent were to consider investing a tiny bit less. It would suffer a slightly increased probability of losing the patent to the challenger and a slightly decreased chance of collecting the patent itself, but would spend a bit less and would receive the flow revenue $R$ stochastically longer. The challenger, by investing a bit less, suffers a slightly increased probability of losing the patent to the incumbent and a slightly decreased probability of collecting the patent for itself; on the other hand, it spends a bit less. Since it *does not* collect any additional current revenue, its marginal benefits due to investing a bit less are lower than those of the incumbent, and hence in equilibrium the challenger invests more than the incumbent. Consider the same question under the assumption of certainty. What happens in the certainty model if the incumbent were to consider investing a tiny bit less? If the incumbent still invests more than the challenger, then the incumbent collects revenues $R$ with probability 1 and suffers no threat of losing the patent to the challenger. If the incumbent was investing less than the challenger, then further reductions have no impact on their profits. The only important case is when the incumbent considers reducing its investment from above that of the challenger to below that of the challenger. This results in the incumbent receiving $R$ for an infinitesimally short additional time, and losing the noninfinitesimal difference between the present values of monopoly profits and Nash-Cournot profits when the challenger patents the new technology. Consequently, the incumbent is always willing to invest more than the challenger when the innovation process is deterministic.

## III. A Model with Parallel Projects

In this section, I report briefly on a model in which firms are allowed to select both the number of parallel projects to undertake, denoted $n_I$ and $n_C$ for incumbent and challenger, respectively, and the rate of investment on each project, $x_I$ and $x_C$, respectively.[3] The conjecture is that the incumbent is more likely than the challenger to patent the innovation. However, it can be shown that both firms will select the same level of investment per project; that is, $x_I^* = x_C^*$. Thus each firm varies its "scale" by choosing the number of (statistically independent) projects. Moreover, it can be shown that if the innovation is drastic and $R > 0$, then $n_I^* < n_C^*$; that is, the incumbent invests in fewer parallel projects than does the challenger. Again by the continuity of these equilibrium strategies in the parameters $R, c$, there exists an open set of innovations which are not drastic, for which the incumbent still operates fewer parallel projects than does the challenger.

Thus the alternative form of the conjecture is also false for an open set of innovations—those which are sufficiently revolutionary. It fails for essentially the same reason as before; the incumbent has a lower marginal incentive to hasten the date of innovation, since it continues to receive the flow profit $R$ until innovation, while the challenger does not.

## IV. Conclusion

It seems clear that the assumption of certainty in the inventive process is not an innocuous one, particularly when one compares the policy implications of these two models. The Gilbert-Newbery model suggests that one ought to be very worried about the development of entrenched monopolies via preemptive patenting. This study suggests that one can reasonably worry far less on this score when the inventive process is stochastic.

It seems reasonable that the degree of cost reduction and the degree of associated uncertainty are related. That is, more drastic innovations may also be subject to greater uncertainty. Thus one can reconcile the Gilbert and Newbery paper with Scherer's observations by suggesting that the certainty model is most appropriate for incremental innovations.

Of course, the models discussed in this paper also rely upon simplifying assumptions. The assumption of constant returns to scale in the output production function is particularly useful, since it allows us to use simple parametric expressions for the post-innovation profit functions. Taken together, this paper and that of Gilbert and Newbery indicate that the influence of monopoly power on the persistence of monopoly is considerably more complicated than either paper taken alone might suggest.

## APPENDIX: PROOF OF PROPOSITION 1

The method of proof is as follows. I first show that there is a finite investment level for each firm such that the firm's best response never exceeds this level, regardless of its rival's investment. Then we can restrict the firm to strategies within this interval without any loss of generality. Next it is shown that the firm's payoff function is single peaked in its own investment level, and reaches this peak on the aforementioned compact interval. Finally, we need to show that this peak is a stationary point and that the payoff function is locally concave at this point. This suffices to define the firm's best response function. A fixed-point argument completes the proof.

---

[3] Since it would seem to add little to the essential intuition of the previous section, the analysis of this more general model is not included here; a technical appendix to this paper which contains the analysis of this more general model is available from the author upon request.

$$\partial V^I/\partial x_I = \big[h'(x_I)\big(\Pi(c)r - R + h(x_C) \times (\Pi(c) - \pi_I(c)) + x_I\big) - B\big]/B^2,$$

where $B = r + h(x_I) + h(x_C)$. Let

$$f_I(x_I, x_C) = h'(x_I)\big[\Pi(c)r - R + h(x_C) \times (\Pi(c) - \pi_I(c)) + x_I\big] - B.$$

Note that $\operatorname{sgn} \partial V^I/\partial x_I = \operatorname{sgn} f_I$. Under the assumption that $h'(0) \geqslant \max\{1/(\Pi(c) - R/r), 1/\pi_C(c)\}$, it can be seen that $f_I(0, x_C) \geqslant 0$ for all $x_C$. Moreover, $f_I(\hat{x}_I(c, R), x_C) \leqslant 0$ for all $x_C$, where

$$\hat{x}_I(c, R) = \min\{x \in [0, \infty) | h'(x) \leqslant \min\{1/(\Pi(c) - \pi_I(c)), 1/(\Pi(c) - R/r)\}\}.$$

This value exists and is finite since $\lim_{x \to \infty} h'(x) = 0$. Since

$$\partial f_I/\partial x_I = h''(x_I)\big[\Pi(c)r - R + h(x_C) \times (\Pi(c) - \pi_I(c)) + x_I\big]$$

is strictly negative, the function $V^I$ is first increasing, eventually peaks, and subsequently declines. Consequently, $V^I$ is single peaked and reaches its peak at or before $\hat{x}_I(c, R)$ for all $x_C$.

The value of $x_I$ which provides the peak is $\phi_I(x_C; c, R)$, the unique best response for the incumbent to $x_C$. Moreover, since $V^I$ is twice differentiable in $(x_I, x_C)$ and continuous in $(c, R)$, and since

$$\partial^2 V^I(\phi_I, x_C)/\partial x_I^2 = h''(\phi_I)\big[\Pi(c)r - R + h(x_C)(\Pi(c) - \pi_I(c)) + \phi_I\big]/B^2 < 0,$$

$\phi_I$ is implicitly defined as a continuously differentiable function of $x_C$ (and a continuous function of $(c, R)$) by the first-order condition $\partial V^I/\partial x_I = 0$. A similar argument establishes the analogous result for $\phi_C$.

Define the composite function $\omega = \phi_I \circ \phi_C : [0, \hat{x}_I(c, R)] \to [0, \hat{x}_I(c, R)]$ (holding $c$ and $R$ fixed). The function $\omega(x; c, R)$ is continuously differentiable in $x$ on a compact, convex and nonempty domain. Hence it has a fixed point $x_I^*(c, R)$ by Brouwer's theorem; that is, there is a point $x_I^*(c, R)$ such that $\omega(x_I^*(c, R); c, R) - x_I^*(c, R) = 0$. Under the assumption that $x_I^*(c, R)$ is not a critical point of $\omega(x; c, R) - x$ (i.e., $\partial\omega(x_I^*(c, R); c, R)/\partial x \neq 1$), there exists a neighborhood of $c$ in which the implicit function $x_I^*(c, R)$ is continuous (see Magnus Hestenes, p. 22, Theorem 7.1). Let $x_C^*(c, R) = \phi_C(x_I^*(c, R); c)$. The strategies $x_I^*(c, R)$ and $x_C^*(c, R)$ constitute a Nash equilibrium, and they are continuous in the parameter $c$.

## REFERENCES

**Craswell, Richard,** "Patents, Sleeping Patents and Entry Deterrence: Comments," in Steven C. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Federal Trade Commission, September 1981, 271–85.

**Dasgupta, Partha and Stiglitz, Joseph,** "Uncertainty, Industrial Structure and the Speed of R and D," *Bell Journal of Economics*, Autumn 1980, *11*, 1–28.

**Gilbert, Richard J.,** "Patents, Sleeping Patents and Entry Deterrence," in Steven C. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Federal Trade Commission, September 1981, 205–69.

______ **and Newbery, David M. G.,** "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

**Hestenes, Magnus R.,** *Calculus of Variations* *and Optimal Control Theory,* Huntington: Robert E. Krieger, 1980.

**Lee, Tom and Wilde, Louis L.,** "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics,* March 1980, *94,* 429–36.

**Loury, Glenn C.,** "Market Structure and Innovation," *Quarterly Journal of Economics,* August 1979, *93,* 395–410.

**Scherer, F. M.,** *Industrial Market Structure and Economic Performance,* 2d ed., Chicago: Rand McNally & Co. 1980.

# Plea Bargaining and Social Welfare

*By* GENE M. GROSSMAN AND MICHAEL L. KATZ*

It is estimated that over 90 percent of convictions in criminal cases in the United States result from a negotiated plea of guilty.[1] It is imperative, therefore, that social scientists and lawyers study the practice of plea bargaining to gain an understanding of the role that it plays in the functioning of the American legal system, and to evaluate its implications for social welfare. The tools of economic analysis are particularly well suited to this endeavor, as the plea bargain is similar in form to a voluntary commercial transaction.

Previous studies of plea bargaining by economists have focused primarily on its conservation of economic resources. For example, William Landes (1971) considers the conditions under which the plea bargain benefits (in expected utility terms) both the prosecutor and the defendant, assuming that trial involves considerable material input from both parties. Richard Adelstein (1978) extends the Landes analysis to incorporate the time-related costs of trial proceedings. These authors conclude, as did Newman, writing from a different perspective, that the more costly in financial terms is a trial conviction, the more economically efficient is the plea bargain.

While the conservation of resources devoted to the legal system undoubtedly is an important objective, this justification of plea bargaining raises some difficult constitutional problems.[2] In this paper, we study two additional benefits derived from plea bargaining that are not subject to these problems, and that have been neglected. First, the plea bargain acts as an *insurance device* for both the innocent defendant and the state.[3] One mandate of the judicial process in the United States is to ensure that the innocent are not wrongly punished. This fact generally has been regarded as an argument against plea bargaining, since the innocent defendants may be led by lenient punishment falsely to plead guilty. Although such erroneous convictions certainly are costly in terms of social welfare, so too are cases in which the trial process produces incorrect findings. When such cases do occur, the penalties imposed will be more severe than the plea bargain, and the social costs will be larger. We argue, therefore, that the constitutional mandate to protect the innocent should be interpreted to imply that the state's interest where an innocent defendant is concerned is identical to that of the individual. This tenet forms the basis for the prosecutor's objective function in our model.

Another mandate of the legal system is to punish the guilty appropriately. Again, lenient plea bargains impose social costs. But so too do cases in which the guilty are acquitted wrongly. A risk-averse society will prefer a sure conviction of the guilty under a range of somewhat more lenient terms to the inherently risky process of litigation.

The second, and perhaps more novel, role of the plea bargain that we emphasize is its role as a *screening device*. The legal system is

---

*Department of Economics, 113 Dickinson Hall, Princeton University, Princeton, NJ 08544. We thank Richard Adelstein, Avinash Dixit, Jean Baldwin Grossman, Carl Shapiro, Steven Shavell, and Marilyn Simon for helpful comments.

[1] See Donald Newman (1966). Thomas McCoy and Michael Mirra (1980) cite J. Kaplan (1977) as obtaining a similar finding.

[2] McCoy and Mirra have argued that the due process clause of the 14th amendment prohibits the state from penalizing an individual for the exercise of a constitutional right, where "penalty" has been interpreted to include the withholding of a reward (for example, a lenient sentence) for the exercise of a right (for example, the right to trial). The authors then argue that the Supreme Court has ruled that a reduction in the state's expenditures is not a sufficient benefit to offset the imposition of a burden on the exercise of a fundamental right.

[3] Landes does incorporate risk-averse defendants. However, he neglects the implications of the possible existence of innocent defendants, and he adopts a prosecutorial utility function that is different from ours.

fundamentally characterized by asymmetric information; the accused know whether they are guilty, while the prosecutor never can be certain about a given defendant's guilt or innocence. In this environment, the plea bargain often can be chosen to serve as a self-selection mechanism. By doing so, it contributes simultaneously to the accuracy of the legal process (by inducing the guilty to identify themselves) and to ensuring that violators of the law are indeed punished.[4]

To highlight these two arguments in favor of plea bargaining, and to show how powerful they potentially are, we will ignore completely the resource cost of the legal system and will significantly simplify the bargaining problem associated with the negotiation of the plea.

The paper is organized as follows. In Section I, we discuss the appropriate specification of the social objective function. We also briefly consider the implications of our model in the case where all defendants actually are guilty. We show that if either the defendant or society is risk averse, then the plea bargain is socially beneficial as an insurance device.

The heart of our analysis is contained in Section II, where we assume that some proportion of the defendants are not guilty, that all defendants are equally risk averse, and that the prosecutor cannot identify the truly innocent.[5] We show that for some parameter values, the optimal plea bargain acts as a screening device, such that only those who are actually guilty so plead, while the innocent choose instead to engage in trial. Furthermore, if the prosecutor also has influence over the penalty imposed upon conviction at trial, the optimal plea bargain always serves this "separating" function.

In Section III, we show that these results may be too sanguine. When the assumption that all defendants are equally risk averse is relaxed, we find that the social usefulness of the plea bargain may be diminished. Intuitively, this is because innocent but highly risk-averse individuals may be behaviorally indistinguishable to the prosecutor from those who are guilty but less risk averse. The state's optimal policy still involves some plea bargaining, but complete separation is no longer attainable. In general, some innocent may be led to plead guilty, while other, truly guilty defendants fail to identify themselves as such.

## I. The Legal Process: A Model

A basic assumption throughout our analysis is that both the defendant and the prosecutor behave rationally, in the sense that each acts to maximize an appropriately defined utility function. The defendant, guilty or innocent, wishes to minimize his punishment. Let $p$ represent the punishment in whatever units are natural, and define the defendant's utility function as $U(p)$. We normalize so that $U(0)=0$, and we assume $U'<0$ and $U''<0$ (i.e., the defendant is risk averse).

In choosing a form for the prosecutor's objective function, we depart significantly from previous studies in two ways. First, we assume that the prosecutor acts entirely in the state's interest, so as to maximize the welfare of society.[6] Second, we argue that the social welfare function must incorporate not only utility losses associated with type II errors (i.e., failure to accurately identify and appropriately punish guilty actors), but also those associated with type I errors (i.e., unjustified conviction of innocent defendants). Indeed, it is the latter source of error that forms the basis for all constitutional requirements of due process in the criminal procedure.

In the case of an innocent party, it seems plausible that the state's interest should coincide with that of the individual. The loss to the individual deriving from erroneous punishment, which we have written as $U(p)$,

---

[4] In the words of McCoy and Mirra, the plea bargain so utilized satisfies the test of "procedural due process."

[5] Presumably, if the prosecutor acting in the state's interest could make this distinction, he would not press charges against the innocent.

[6] If the effort of the prosecutor is not perfectly observable by his superiors, then a principal-agent problem may result (see, for example, Bengt Holstrom, 1979, and Steven Shavell, 1979). The prosecutor's behavior then would depend on the incentive structure that he faced. This interesting possibility is not considered here.

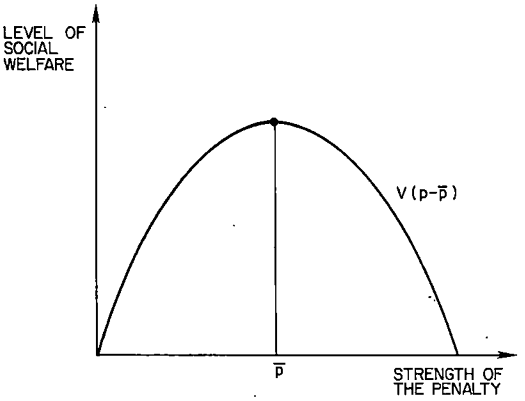


FIGURE 1. SOCIAL WELFARE DERIVED FROM PENALIZING THE GUILTY

is a loss from society's perspective as well.[7] Naturally, where a guilty defendant is concerned, the interests of the state diverge from those of the individual. Here we adopt the construct of "a penalty that fits the crime"; that is, a single measurable penalty $\bar{p}$ that society deems most appropriate given the statutorily defined misconduct.[8] Social welfare then is a function of the deviation of the actual penalty from the most-appropriate penalty, given by $V(p-\bar{p})$. This function, which is depicted in Figure 1, achieves its maximum value at $p=\bar{p}$. Society is assumed to be risk averse, that is, $V''<0$.

We assume that the prosecutor does not know *ex ante* whether a particular individual is guilty or innocent, but that the prosecutor can estimate from experience the fraction, $\lambda$, of guilty defendants in the total population of indictees. When all defendants receive penalty $p$, expected social welfare $W$ can be written as

$$(1)\quad W=\lambda V(p-\bar{p})+(1-\lambda)U(p).$$

Consider now the legal process, which involves either a pretrial settlement or a trial proceeding. In reality, plea bargaining is just that: a negotation that may involve give and take, the result of which will depend on the relative strengths of the bargaining positions and the abilities of the attorneys involved. In order to concentrate on other aspects, we simplify this process considerably by assuming that the prosecutor makes a single, take-it-or-leave-it offer. Some justification for this assumption derives from the fact that the prosecutor is involved in the bargaining game a large number of times, and in this game a reputation for toughness is a valuable asset. The defendant, by contrast, only engages in the bargaining game infrequently.[9] Reality is further complicated by the fact that the prosecutor does not have legal authority to determine the final sentence. Although the prosecutor can influence the penalty significantly by recommendation of a lenient sentence, or by a reduction in charges, ultimate discretion lies solely with the judge. Nonetheless, we assume here that the prosecutor can offer a guaranteed penalty of $p_b$.

The defendant may accept this offer, or opt for a trial. Each trial is fully characterized by a probability of conviction, $\theta$, and a penalty imposed following a finding of guilty, $p_t$. While remaining agnostic on the accuracy of the courtroom procedure, we do assume that a trial does have some incremental value in determining the truth. Operationally, this assumption implies that the probability that a truly guilty defendant will be found guilty, $\theta_{gg}$, exceeds the probability that an actually innocent defendant will be convicted, $\theta_{ig}$.[10] We further assume that the prosecutor and defendant have identical assessments of these subjective probabilities.[11] Of course, the defendant is aware of his true state, that is, whether he is guilty or innocent.

---

[7] One might even argue further that the disutility to the state derived from erroneous convictions of innocent defendants exceeds the disutility to the individual involved for two reasons: first, the state may have a concern for the integrity of the system; second, punishment may involve a direct resource cost. The inclusion of such considerations would not alter our analysis in any fundamental way.

[8] In some cases, for example when damages are measurable in financial terms, this construct may be conceptually well defined. In many other cases, the construct will be less well defined and will reflect prevailing views in the field of penology.

[9] David Kreps and Robert Wilson (1980) employ a similar argument in their analysis of a monopolist who is confronted in successive periods by a set of potential entrants into its industry.

[10] Our notation is as follows: the first subscript refers to the true state of the defendant; the second refers to the finding at trial. It follows that $\theta_{gi}=1-\theta_{gg}$, $\theta_{ii}=1-\theta_{ig}$, and $\theta_{ii}>\theta_{gi}$. The subscripts $g$ and $i$ refer to guilty and innocent, respectively.

Let us define the *certainty equivalent* offer, $c_j$ for $j = g, i$, as the guaranteed-penalty offer that leaves an individual who is actually of type $j$ just indifferent between exercising the right to trial and not. Then, by definition,

$$U(c_j) = \theta_{jg} U(p_t), \qquad j = g, i. \tag{2}$$

Finally, we entertain two possible assumptions regarding the value of $p_t$, the penalty imposed upon conviction. First, we may suppose that societal notions of "fairness" will dictate that $p_t = \bar{p}$, the most-appropriate penalty for the crime. However, it can be shown that this penalty is not, in general, socially optimal. We consider therefore the alternative possibility that the posttrial sentence is chosen by the prosecutor along with the plea bargain offer to maximize social welfare. This may be plausible if the prosecutor has considerable discretion over which charges to press, as is common in cases where several statutes have been violated.

In the remainder of this section, we examine the role of the plea bargain as insurance. To highlight the insurance role as one separate from that of sorting, we consider the model under the assumption that all defendants are guilty.[12] The forces identified in this section remain present in the more realistic cases considered in the main part of our analysis, where there are both innocent and guilty defendants.

Suppose now that $\lambda = 1$ (i.e., all indictees are guilty). We will show that, given any *arbitrary* trial penalty, there is a plea bargain that yields a higher level of social welfare than is attained by trial.

The expected utility from trial is

$$W_t = \theta_{gg} V(\bar{p} - p_t) + (1 - \theta_{gg}) V(\bar{p}).$$

Since society is risk averse (i.e., $V$ is a concave function),

$$W_t < V\left[\theta_{gg}(\bar{p} - p_t) + (1 - \theta_{gg})\bar{p}\right]$$
$$= V(\bar{p} - \theta_{gg} p_t).$$

Now suppose the prosecutor offers a plea bargain with penalty $p_b$. The offer will be accepted by defendants if and only if $p_b \leqslant c_g$. There are two cases to consider.

(*i*) If $p_t$ is set such that $c_g < \bar{p}$, let the prosecutor choose $p_b = c_g$. By the concavity of $U$, $U(c_g) = \theta_{gg} U(p_t) < U(\theta_{gg} p_t)$. Hence $c_g > \theta_{gg} p_t$. That is, the certainty equivalent penalty is larger than the expected penalty obtained through trial. The expected social welfare attained by this plea bargain is

$$W_b = V(\bar{p} - c_g) > V(\bar{p} - \theta_{gg} p_t) > W_t. \tag{3}$$

(*ii*) If the trial penalty is so large that the certainty equivalent offer exceeds the most-appropriate penalty, then the prosecutor can offer the penalty $\bar{p}$ as a plea bargain and, with all defendants actually guilty, this outcome is the *optimum optimorum*.

To summarize, when all defendants are guilty, the plea bargain can be used to achieve a level of social welfare greater than that attainable through trial. This is because with full information available to the prosecutor, the trial can only add randomness to the legal system, which both the defendant and society generally would wish to avoid. Note that the plea bargain considered above is the socially optimal offer. A higher offer either would not be accepted or would exceed the most appropriate penalty, while a lower one would needlessly provide surplus to the guilty party.

---

[11] This assumption is based on recognition of the constitutional requirement for the state's provision of information. The Supreme Court has ruled in *U.S. v. Agurs*, 427 U.S. 97 (1976) and *Brady v. Maryland*, 373 U.S. 83 (1963) that the defendant must be fully informed of all evidence available to the state, and of the facts needed by the state to establish the case against him, prior to the arraignment and choice of a plea. It is illegal for the prosecutor to conceal or falsify evidence in plea negotiations, as it would be to do so in trial.

[12] When all defendants are assumed to be guilty, our model is essentially a special case of the one studied by Landes.

## II. The Plea Bargain as a Screening Device

Now let us suppose, more realistically, that only a fraction of those who are indicted are in fact culpable. The prosecutor can estimate $\lambda$, but cannot identify those who are actually guilty. We assume in this section, that all individuals, guilty or innocent, are equally risk averse. This turns out to be an important assumption, as we will demonstrate in Section III.

Consider once again the settlement offer $p_b = c_g$.[13] We call this offer the "best separating plea bargain." It is "separating," because only the guilty will accept such an offer; the innocent will opt for trial if faced with any plea bargain offer greater than $c_i$, and $c_g > c_i$. Among the set of separating offers, namely those between $c_i$ and $c_g$, the offer $p_b = c_g$ is "best," because it extracts a penalty from the guilty that is closest to the social norm.

Let $W_s$ be expected social welfare achieved by the best-separating plea bargain, that is,

$$(4) \quad W_s = \lambda V(\bar{p} - c_g) + (1-\lambda)\theta_{ig} U(p_t).$$

Compare this level of welfare with expected social welfare level when all defendants go to trial,

$$(5) \quad W_t = \lambda\theta_{gg} V(\bar{p} - p_t) + \lambda(1-\theta_{gg}) \times V(\bar{p}) + (1-\lambda)\theta_{ig} U(p_t).$$

Clearly, $W_s > W_t$ if and only if

$$V(\bar{p} - c_g) > \theta_{gg} V(\bar{p} - p_t) + (1-\theta_{gg}) V(\bar{p}).$$

But this comparison is the same one made earlier to show that the plea bargain dominates the trial procedure when all defendants are guilty. We thus have established the following proposition.

PROPOSITION 1: *If all defendants are equally risk averse, the best separating plea bargain dominates no plea bargaining (i.e., court proceedings for all defendants).*

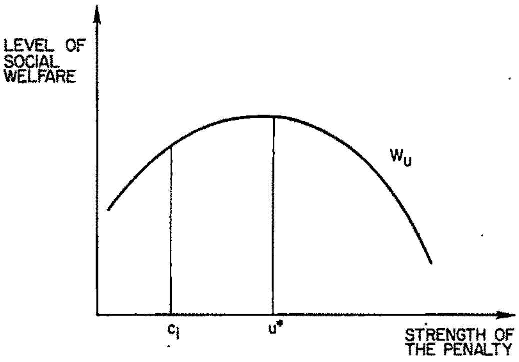


FIGURE 2. SOCIAL WELFARE DERIVED FROM A UNIFORM PENALTY

An immediate implication of this proposition is that the *optimal* plea bargain always extracts a guilty plea from all the truly guilty defendants.

Is $p_b = c_g$ the optimal policy for the prosecutor? An alternative policy is to offer a plea bargain that is accepted by both innocent and guilty defendants. Such an offer is said to be a "pooling" or "nonseparating" policy. Let $u$ be a uniform punishment that is imposed on innocent and guilty alike. Social welfare associated with $u$ is $W_u$, and given by

$$(6) \quad W_u = \lambda V(\bar{p} - u) + (1-\lambda) U(u).$$

Among the set of uniform punishments, the best (which we shall denote $u^*$) is the one that maximizes (6). The first order condition for this maximization (assuming $u^* > 0$) is

$$(7) \quad U'(u^*)/V'(\bar{p} - u^*) = \lambda/(1-\lambda).$$

The second-order sufficiency condition is satisfied, because $W_u'' = \lambda V'' + (1-\lambda)U'' < 0$. Social welfare as a function of $u$ is depicted in Figure 2.

If the solution to (7) yields $u^* > c_i$ (as drawn in Figure 2), then an offer of $u^*$ as a plea bargain would be declined by innocent defendants. Thus, it would not be feasible as a pooling plea bargain. Among feasible pooling offers (i.e., those less than or equal to $c_i$), we see from Figure 2 that $p_b = c_i$ is best. But this plea bargain is clearly inferior to the best separating offer, which yields the same utility for innocent defendants (recall that $c_i$ is the certainty equivalent) and provides a more appropriate penalty for the guilty.

---

[13] In order to simplify the exposition, we henceforth assume that $c_g \leqslant \bar{p}$. We otherwise could proceed as above with $p_b = \min(c_g, \bar{p})$, and the argument could be made rigorously for the most general case.

Suppose, instead, that the solution to (7) yields $u^* \in (0, c_i)$, and consider the case in which the trial penalty is set equal to the most-appropriate penalty, that is, $p_t = \bar{p}$. Then since $u^*$ *is* now feasible as a pooling plea bargain, it is the best among this set of policies and results in welfare of $W_u = \lambda V(\bar{p} - u^*) + (1 - \lambda) U(u^*)$. We cannot rule out the possibility that this level of welfare is larger than that achieved by the separating plea bargain, $W_s = \lambda V(\bar{p} - c_g) + (1 - \lambda)\theta_{ig} U(\bar{p})$.

Two offsetting forces are at work in this comparison. The separating plea bargain extracts a larger and socially more appropriate penalty from the guilty than does the pooling plea bargain. But the separating plea bargain also yields a lower level of utility to the innocent defendants, who face the possibility of incorrect conviction through trial and the concomitant high penalty. The fact that defendants self-select implies that innocent defendants must be better off under the pooling plea bargain, since the option of going to trial remains open when the pooling offer is made. Thus, the greater is the social loss associated with incorrect convictions, the greater the proportion of truly innocent persons among defendants, and the less effective is the trial process in identifying the true state of the defendant, the more likely it is that the pooling plea bargain is superior to the separating plea bargain.[14] We summarize our findings as follows.

PROPOSITION 2: *Assume all defendants are equally risk averse, and that* $p_t = \bar{p}$. *Let* $u^*$ *be the best uniform penalty, irrespective of feasibility as a plea bargain. Then* (i) *if* $u^* \geqslant c_i$ *the optimal plea bargain is the best separating offer,* $c_g$. (ii) *If* $u^* \in (0, c_i)$ *the optimal plea bargain is either the best separating offer or the pooling offer,* $u^*$.

The prosecutor might not, however, take the trial penalty as given, as we have assumed above. He may instead be able to influence the sentence imposed upon conviction at trial through variation in the charges that are pressed. Let us suppose then, for purposes of comparison, that the prosecutor is capable of determining $p_t$, and does so optimally. This extra flexibility proves to be qualitatively important for the determination of the prosecutor's optimal policy.

PROPOSITION 3: *If all defendants are equally risk averse, and if the prosecutor chooses both* $p_b$ *and* $p_t$, *then social welfare is maximized by a separating plea bargain.*

PROOF:

The best separating policy solves

$$W_s = \max_{p_t, p_b} \lambda V(\bar{p} - p_b) + (1 - \lambda)\theta_{ig} U(p_t) \tag{8}$$

subject to

$$\theta_{gg} U(p_t) \leqslant U(p_b) < \theta_{ig} U(p_t).$$

Since $p_b < \bar{p}$, it is best to extract as large a penalty as possible from the guilty. Therefore, the constraint can be written $\theta_{gg} U(p_t) = U(p_b) < \theta_{ig} U(p_t)$. Substitution yields the unconstrainted maximization

$$W_s = \max_{p_b} \lambda V(\bar{p} - p_b) + (1 - \lambda)\theta_{ig} U(p_b)/\theta_{gg}. \tag{9}$$

The best pooling policy solves

$$W_u = \max_{p_b} \lambda V(\bar{p} - p_b) + (1 - \lambda) U(p_b) \tag{10}$$

subject to $\quad U(p_b) \leqslant \theta_{ig} U(p_t).$

Since $U(p_b) \leqslant 0$ and $\theta_{ig}/\theta_{gg} < 1$, it follows that $W_s \geqslant W_u$.

---

[14] To verify the second claim, note that

$$d(W_s - W_u)/d\lambda = [V(\bar{p} - c_g) - V(\bar{p} - u^*)]$$
$$+ [U(u^*) - \theta_{ig} U(p_t)] \geqq [V(\bar{p} - c_g) - V(\bar{p} - u^*)]$$
$$+ [U(c_i) - \theta_{ig} U(p_t)] = V(\bar{p} - c_g) - V(\bar{p} - u^*) > 0.$$

The intuition for this result is as follows. Suppose we begin with any pooling policy, so that both innocent and guilty accept the prosecutor's settlement offer. Now consider a lowering of the trial penalty $p_t$ by the prosecutor. Initially this has no effect, so long as neither party opts for trial. A critical $p_t$ is eventually reached such that the *innocent* will choose trial over the existing plea offer (recall that $c_i < c_g$). Further reduction in $p_t$ within some range strictly raises social welfare, because it raises the utility of the innocent without affecting the penalty imposed on the guilty. We can always construct in this way a separating policy that is superior to any given pooling policy.

To summarize the conclusions of this section, we have shown that the plea bargain is a valuable social policy when the legal system is characterized by imperfect information. Its usefulness derives from its role as an insurance device, and its potential role as a screening device. When trial sentences are outside the prosecutor's control it may be socially optimal to set the plea offer so that it induces guilty defendants to reveal their guilt by self-selecting the plea offer. Used this way, the plea bargain contributes to the procedural accuracy of the legal system. If the prosecutor can set the trial penalty as well as the penalty for a guilty plea, then the optimal plea bargain is always used to sort the guilty, and innocent defendants are never led to confess falsely. Under these conditions, the case for the plea bargain is strongest. These optimistic results are, however, all premised on the assumption that all defendants are equally risk averse. We must now consider the implications of relaxing this assumption.

## III. The Plea Bargain when Defendants are not Equally Risk Averse

The attractiveness of the plea bargain as a screening device above arises because the prosecutor can perfectly associate actual guilt with a high certainty-equivalent penalty. As long as the court system has some incremental value in ascertaining truth, there will exist a range of plea bargains that only guilty parties will accept. However, such a range may not exist if various defendants differ in their degree of risk aversion. A highly risk-averse but innocent defendant is indistinguishable to the prosecutor from a less risk-averse, guilty defendant, if they have equal certainty-equivalent penalties.[15] If there exists a class of less risk-averse, guilty defendants with a higher certainty equivalent penalty than that of a class of more risk-averse, innocent defendants, then there can be no plea bargain that perfectly separates the innocent from the guilty. We will find that, in general, many plea bargain offers have the characteristic that some guilty elect to go to trial, while some innocent falsely plead guilty. The optimal social policy does still involve some plea bargaining, but the scope of plea bargaining as a screening device may be significantly limited.

To illustrate these points, we consider the prosecutor's policy problem when there are two risk-aversion classes. The subscripts $h$ and $l$ refer to high and low risk aversion, respectively, and we assume $c_{gh} > c_{ih} > c_{gl} > c_{il}$.[16] We show first that, even with multiple risk-aversion classes and trial penalties that are outside the prosecutor's control, the prosecutor's optimal policy always involves some plea bargaining.

PROPOSITION 4: *With two risk-aversion classes, the plea bargain offer* $p_b = c_{gh}$ *dominates no plea bargaining (i.e., court proceedings for all defendants).*

Given that $c_{gh} > c_{ih} > c_{gl} > c_{il}$, the plea offer $p_b = c_{gh}$ will be accepted only by the more risk-averse, guilty defendants. The treatment of all other defendants types is the same as under no plea bargaining. Thus, to ascertain the social desirability of this plea bargain, we need compare only the levels of welfare obtained by sending the more risk-averse, guilty defendants to court to that achieved by offering them the plea bargain $p_b = c_{gh}$. This comparison is identical to that in Section I, where we analyzed the role of plea bargaining as insurance. Here, as there, the plea bargain can be used to insure society against acquittal of the most risk-averse, guilty defendants.

---

[15] This observational equivalence is analogous to that studied in another context by Joseph Stiglitz (1977), who found that the ability of a monopolistic insurance company to screen high-risk agents from low-risk agents is diminished if the agents differ not only in their probability of accident, but also in their degree of risk aversion.

[16] The alternative possible configuration, $c_{gh} > c_{gl} > c_{ih} > c_{il}$, is less interesting because a plea bargain offer set in the range $(c_{ih}, c_{gl})$ acts to separate the guilty from the innocent perfectly. Since the analysis of this case is formally identical to that studied above, we will not pursue it further here.

For the case where the trial penalty is equal to the social norm (i.e., $p_t = \bar{p}$), $p_b = c_{gh}$ is one candidate for the optimal plea bargain. But there are now several other candidates to consider. We can rule out immediately all plea bargains in the range $c_{gh} > p_b \geqq c_{ih}$. Offers in this range extract a smaller penalty from the more risk-averse, guilty defendants than $p_b = c_{gh}$, and these offers fail to raise the utility of either innocent group. But a plea bargain in the range $c_{ih} > p_b > c_{gl}$ raises the utility of the more risk-averse, innocent defendants, albeit by causing them to falsely plead guilty. If this utility increase is large enough, it might outweigh the social loss caused by lenient treatment of the more risk-averse, guilty defendants, and the optimal plea bargain might lie in this range. Plea bargains in the range $c_{gl} \geqq p_b > c_{il}$ cannot be ruled out either. These offers will be accepted by all but the less risk-averse, innocent defendants, and yield surplus to both groups of guilty defendants. Nonetheless, an offer in this range can be optimal if the utility gain on the more risk-averse, innocent group is large enough (for example, if this group is very large). Finally, the optimal plea bargain may also fall in the range $c_{il} \geqq p \geqq 0$, in which case all defendants plead guilty. If the social loss associated with type I errors of the legal system (i.e., wrongly punishing the innocent) is large enough, the system is rendered virtually inoperative, with lenient treatment provided to all who are indicted.

It appears, then, that the optimal plea bargain under exogenous trial penalty may involve lenient treatment of the guilty, false pleas by the innocent, failure to induce self-selection by some guilty defendants, or a combination of these outcomes. It is clear that the effectiveness of the plea bargain as a sorting device in this environment is considerably less than it would be in the case where all defendants were equally risk averse.

The pessimism of this result is somewhat mitigated if the prosecutor can select the trial penalty. The following proposition deals with this case.

PROPOSITION 5: *With two risk-aversion classes, if the prosecutor sets the trial penalty optimally, then the optimal plea bargain always is accepted by the more risk-averse, guilty defendants and never accepted by the less risk-averse, innocent defendants.*

The ability to set the trial penalty provides the prosecutor with an additional instrument to raise the utility of the innocent. Starting from any settlement offer and trial penalty combination such that all defendants plead guilty, social welfare can be raised by lowering the penalty at least to the point where the less risk-averse, innocent defendants opt for trial proceedings. The prosecutor thus raises the utility of this innocent group, without affecting the penalty imposed on either group of guilty defendants.

Under the conditions of Proposition 5, the plea bargain does play a partial screening role. One group of guilty defendants always is induced to identify itself as such, while one group of truly innocent indeed pleads not guilty. The optimal plea bargain will, however, either fail to induce self-selection of the less risk-averse guilty, or will extract a false guilty plea from the more risk-averse innocent. This is an inevitable consequence of the prosecutor's imperfect information that arises whenever the certainty equivalent penalty of an innocent group exceeds that of a guilty group.

We conclude by noting that with a large number of groups of defendants with varying degrees of risk aversion, it may be optimal to use the plea bargain to self-select only a small fraction of the guilty. Nonetheless, a plea bargain offer set somewhat close to the trial penalty will always be welfare improving if it causes the risk-averse guilty to plead guilty, thereby insuring society against the risk of an erroneous trial outcome for these indictees.

## IV. Conclusions

In this paper we have shown that a welfare-theoretic argument in favor of plea bargaining need not be based on a consideration of the resource cost of trial proceedings. The plea bargain can serve equally or more important functions as an insurance device and a screening device. In the former role, it insures society against erroneous outcomes in the courtroom. In the latter role, it sorts the guilty and innocent, in realistic situations where the prosecutor cannot distinguish the falsely accused from the truly guilty. As a self-selection mechanism, it contributes to the accuracy of the legal procedure, thereby aiding in the satisfaction of the constitutional mandate for due process.

We found that the screening role of the plea bargain is most effective when all defendants are equally risk averse and when the prosecutor can influence the sentence imposed on those who are convicted. In this case, socially optimal policy always involves a guilty plea by the truly guilty, and never involves false pleas by the innocent. When defendants may differ in their aversion to risk, the prosecutor may be unable to distinguish an innocent, but highly risk-averse defendant from a guilty, but less risk-averse one. Although some plea bargaining should still be undertaken in such a world, its scope for raising the level of welfare achievable by the legal system over that associated with a trial-for-all system may be limited.

## REFERENCES

**Adelstein, Richard P.**, "The Plea Bargain in Theory: A Behavioral Model of the Negotiated Guilty Plea," *Southern Economic Journal*, January 1978, *44*, 488–503.

**Holstrom, Bengt**, "Moral Hazard and Observability," *Bell Journal of Economics*, Spring 1979, *10*, 74–91.

**Kaplan, J.**, "American Merchandising and the Guilty Plea: Replacing the Bazaar with the Department Store," *American Journal of Criminal Law*, 1977, *5*, 215–24.

**Kreps, David M. and Wilson, Robert**, "On the Chain Store Paradox and Predation: Reputation for Toughness," Technical Report No. 317, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1980.

**Landes, William M.**, "An Economic Analysis of the Courts," *Journal of Law and Economics*, April 1971, *14*, 61–108.

**McCoy, Thomas R. and Mirra, Michael J.**, "Plea Bargaining as Due Process in Determining Guilt," *Stanford Law Review*, May 1980, *32*, 887–941.

**Newman, Donald J.**, *Conviction: Determination of Guilt or Innocence Without Trial*, Boston: Little, Brown and Co., 1966.

**Shavell, Steven**, "Risk Sharing and Incentives in the Principal and Agent Relationship," *Bell Journal of Economics*, Spring 1979, *10*, 55–73.

**Stiglitz, Joseph E.**, "Monopoly, Non-linear Pricing and Imperfect Information: The Insurance Market," *Review of Economic Studies*, October 1977, *44*, 407–30.

# Government Irrelevance Results: A Simple Exposition

*By* JOHN BRYANT*

Recently, a rash of irrelevance of government results has appeared in the macroeconomics literature (see, for example, Robert Barro, 1979; Thomas Finn, 1982; Christian Gilles and Colin Lawrence, 1981; Joseph Stiglitz, 1981; Neil Wallace, 1981). To the novice, particularly one trained in the IS-LM paradigm, these results appear mysterious, if not down right ridiculous. These irrelevance results have typically been derived in models of some generality, and quite appropriately so. However, this generality has served, it seems, to obscure the very simple nature of the essential insights behind the government irrelevance results. It is the goal of this paper to lay bare these simple essential insights in the context of the simplest model I can craft to this purpose. I wish to stress that it is my purpose to clearly illustrate government irrelevance results, not to trivialize them.

The government irrelevance results involve the expenditure for and financing of transfer payments, government investment, and public goods. Because models of fiat money are not simple, nor the presence of fiat money important to our purpose, my model is a nonmonetary model.

## I. The Basic Model

The economy consists of a large number of identical individuals. Time is discrete and is divided into two periods, and individuals live for both periods. There is a single consumption good which can be consumed in both periods. Each individual is endowed with $L$ units of the consumption good in period 1 and 0 units in period 2. Any endowment unconsumed in the first period can be planted to yield $x$ units of consumption good in the second period per unit planted, and in that proportion for fractional units. The common, nicely shaped, utility function is $U(C_1, C_2)$, where $C_1$ and $C_2$ are the individual's consumption in periods 1 and 2, respectively. Let $I$ be the amount of first-period consumption planted by the representative individual. As the individual throws nothing away his or her first-period problem can be written

$$(1) \qquad \max_I U[L-I, xI].$$

This is solved with $I$ satisfying

$$(2) \qquad -U_1[L-I, xI] + xU_2[L-I, xI] = 0.$$

## II. Transfer Payments

To consider the expenditure and financing of transfer payments, we have to add an outside entity, the government, to the basic model.

The first irrelevance result to be produced is that transfer payments financed by taxes are inconsequential. This result we obtain almost for free. Suppose the government hands each individual a lump sum transfer payment in the consumption good in the amount of $G$ in the first period and takes away from each individual a lump sum tax in the consumption good in the amount $T$ in the first period. The individual's first period problem then becomes

$$(3) \qquad \max_I U[L+G-T-I, xI].$$

Assuming the government operates costlessly, the government's budget constraint is $G = T$, thus $G$ and $T$ cancel and (3) replicates (1).

The second irrelevance result produced is that whether transfer payments are financed by taxes or bond issue is inconsequential. Suppose the government finances the above transfer payments by sale of bonds. The government sells, for first-period consumption good in the amount $B$ per person, promises to a total of $(1+r)B$ units of second-period consumption good per person. The individual treats $r$, which is to be determined, as a given, and chooses $B$. The government then services its debt issue in the second period with an equal lump sum tax of $T$ per person, and the individual takes $T$ as given. The individual's first-period problem then can be written

---

*Fox Associate Professor of Economics, Department of Economics, Rice University, Houston, TX 77001.

$$\max_{I,B} U[L+G-B-I, -T+(1+r)B+xI]. \tag{4}$$

This is solved with $I$ and $B$ satisfying

$$-U_1[L+G-B-I, -T+(1+r)B+xI] + xU_2[L+G-B-I, -T+(1+r)B+xI]=0; \tag{5}$$

$$-U_1[L+G-B-I, -T+(1+r)B+xI] + (1+r)U_2[L+G-B-I, -T+(1+r)B+xI]=0. \tag{6}$$

Assuming the government operates costlessly, the first-period government budget constraint is $G=B$ and the second-period government budget constraint is $(1+r)B=T$. These terms cancel and (5) replicates (2). With $x=(1+r)$, (6) replicates (5) and we are back to our original solution, but with $B=G$. Private saving has risen exactly in the amount of the government transfer payment and bond issue, the fruits of which additional saving exactly cover the second-period tax.

### III. Government Investment

The third irrelevance result to be produced is that government investment financed by taxes is inconsequential. Suppose the government plants the consumption good in the amount $G_1$ per person in the first period and takes away from each individual a lump sum tax in the consumption good in the amount $T$ in the first period. The government then in the second period hands each individual a lump sum transfer payment in the consumption good in the amount $G_2$. The individual's first-period problem then becomes

$$\max_{I} U[L-T-I, G_2+xI]. \tag{7}$$

This is solved with $I$ satisfying

$$-U_1[L-T-I, G_2+xI] + xU_2[L-T-I, G_2+xI]=0. \tag{8}$$

Assuming the government operates costlessly, the first-period government budget constraint is $G_1=T$ and the second-period government budget constraint is $G_2=xG_1$, hence $G_2=xT$. Let $\bar{I}$ satisfy (2). Substituting $G_2=xT$ into (8), I conclude that $I=\bar{I}-T=\bar{I}-G_1$ satisfies (8), since then (8) replicates (2). Private investment has fallen exactly in the amount of the government tax and investment, the lost fruits of that private investment exactly being offset by the second-period transfer payment, the fruits of the government investment.

Notice that my irrelevance result depends upon the assumption that $G_1<\bar{I}$. Negative planting is not feasible. If the government plants more than the private sector otherwise would plant, that must increase total planting. However, if, instead of planting, the government bought future contracts for commodity, the private market would short the market, and the irrelevance result goes through.

The fourth irrelevance result to be produced is that whether government investment is financed by taxes or bond issue is inconsequential. Suppose the government finances the above investment by bond issue in the first period, and services its debt issue in the second period by lump sum taxes. The individual's first-period problem then can be written

$$\max_{I,B} U[L-B-I, G_2-T+(1+r)B+xI]. \tag{9}$$

This is solved with $I$ and $B$ satisfying

$$\text{(10)}\quad -U_1[L-B-I,G_2-T+(1+r)B+xI] + xU_2[L-B-I,G_2-T +(1+r)B+xI]=0;$$

$$\text{(11)}\quad -U_1[L-B-I,G_2-T+(1+r)B+xI] +(1+r)U_2[L-B-I,G_2-T +(1+r)B+xI]=0.$$

Assuming the government operates costlessly, the first-period government budget constraint is $G_1=B$ and the second-period government budget constraint is $G_2+(1+r)B=T+xG_1$, hence $G_2+(1+r)B=T+xB$. Let $\bar{I}$ satisfy (2) as above. Substituting $G_2-T+(1+r)B=xB$ into (10), I conclude that $I=\bar{I}-B=\bar{I}-G_1$ satisfies (10), since (10) then replicates (2). With $x=(1+r)$, (11) replicates (10), and we are back to our original solution, but with (once again) government investment substituting for private.

## IV. Public Good

The government's expenditure on a public good is not, of course, inconsequential. Therefore I now turn to the inconsequentiality of the financing of public good expenditure. For the final result, I will produce a weak inconsequentiality result on government public good expenditure.

The fifth irrelevance result is that whether government provision of a public good is financed by taxes or bond issue is inconsequential. To treat public good expenditure, we must alter the basic model somewhat. There is a "third" good, the public good. The government, and only the government, can transform the consumption good in the first period into the public good, one for one. Let $G$ be the per capita expenditure on public good. The individual's utility function is $U(C_1,C_2,G)$. If the government were a benevolent costlessly operating planner its first period problem could be written

$$\text{(12)}\quad \max_{I,G} U[L-G-I,xI,G].$$

This is solved with $I$ and $G$ satisfying

$$\text{(13)}\quad -U_1[L-G-I,xI,G] + xU_2[L-G-I,xI,G]=0;$$

$$\text{(14)}\quad -U_1[L-G-I,xI,G] + U_3[L-G-I,xI,G]=0.$$

Suppose the government produces the public good in the amount $G$ per person in the first period and takes away from each individual a lump sum tax in the consumption good in the amount $T$ in the first period. The individual's first-period problem can be written

$$\text{(15)}\quad \max_{I} U[L-T-I,xI,G].$$

This is solved by $I$ satisfying

$$\text{(16)}\quad -U_1[L-T-I,xI,G] + xU_2[L-T-I,xI,G]=0.$$

Assuming the government operates costlessly, the government's budget constraint is $G=T$. Let $\bar{I}$ and $\bar{G}$ satisfy (13) and (14). With $G=\bar{G}$ and substituting $G=T$ into (16), (16) then replicates (13), so $I=\bar{I}$ satisfies (16). Hence we are back to the planner's solution.

Now suppose the government finances the above public good expenditure by bond issue in the first period, and services its debt issue in the second period by lump sum taxes. The individual's first-period problem then can be written

$$\text{(17)}\quad \max_{I,B} U[L-B-I, -T+(1+r)B+xI,G].$$

This is solved with $I$ and $B$ satisfying .

$$(18)\quad -U_1[L-B-I,-T+(1+r)B+xI,G] + xU_2[L-B-I,-T +(1+r)B+xI,G]=0;$$

$$(19)\quad -U_1[L-B-I,-T+(1+r)B+xI,G] +(1+r)U_2[L-B-I,-T +(1+r)B+xI,G]=0.$$

Assuming the government operates costlessly, the first-period government budget constraint is $G=B$ and the second-period government budget constraint is $(1+r)B=T$. With $G=\bar{G}$ and substituting $G=B$ and $(1+r)B=T$ into (18), (18) then replicates (13), so $I=\bar{I}$ satisfies (18). With $x=(1+r)$ (19) replicates (18), and we are back to the planner's solution.

The sixth, and final, irrelevance result is that, if utility functions are separable in government public good expenditure, allocation of that expenditure is inconsequential for private sector activity. In the case of the separable utility function, $U(C_1,C_2,G)$ can be written as $V(h(C_1,C_2),G)$. Then (16), for example, can be written

$$(20)\quad -V_1h_1(L-T-I,xI) +xV_1h_2(L-T-I,xI)=0.$$

Cancelling the function $V_1$, we get

$$(21)\quad -h_1(L-T-I,xI) +xh_2(L-T-I,xI)=0.$$

Notice that $G$ does not enter this expression (21), only $T$. If the government, instead of producing public good in the amount $G=T$, simply threw the tax proceeds away, (21) would be unaffected. While the individual's utility would be lowered by this profligacy, his or her decision on planting, $I$, would be unaffected. Given separability, private sector activity is uninfluenced, for example, by the allocation of government expenditure between park creation and park destruction. This alone of my irrelevance results should seem familiar to IS-LM students.

## V. Concluding Comments

One should not conclude from the above exercise either that government decisions are inconsequential or that such models are useless. Rather one should conclude that in so far as government decisions matter, they do so for reasons left out of these models. My simple example may well have missed an essential insight of the government irrelevancy literature. However, I do not think so.

## REFERENCES

**Barro, Robert J.**, "On the Determination of the Public Debt," *Journal of Political Economy*, October 1979, *87*, 940–71.

**Finn, Thomas J.**, "Wallace's Irrelevance Proposition for Open-Market Transactions," Wayne State University, 1982.

**Gilles, Christian and Lawrence, Colin**, "A Strong Ricardian Proposition: The Equivalence of Taxes, Bonds and Money Financing in Public Sector Activity," Columbia University, 1981.

**Stiglitz, Joseph**, "On the Relevance or Irrelevance of Public Financial Policy," Princeton University, 1981.

**Wallace, Neil**, "A Modigliani-Miller Theorem for Open-Market Operations," *American Economic Review*, June 1981, *71*, 267–74.

# Expectations, Taxes, and Interest: The Search for the Darby Effect

*By* ROBERT AYANIAN*

The Darby hypothesis states that, *ceteris paribus*, interest rates will rise during an inflation by an amount which exceeds expected inflation in order to compensate lenders both for their expected loss of capital and for the taxation of interest income.[1] The hypothesis is distinguished both by its plausibility and by an almost complete lack of firm statistical support, a statistical condition which this paper seeks to mitigate.

## I. The Issue

Under the Darby hypothesis, the traditional Fisher equation is

$$i = r + \Pi, \tag{1}$$

in which $i$ is the nominal interest rate, $r$ the expected real interest rate, and $\Pi$ the expected inflation rate, must be amended to

$$(1-T)i = r' + \Pi, \tag{2}$$

or

$$i = r'/(1-T) + \Pi/(1-T), \tag{3}$$

where $T$ is the marginal lenders' marginal tax rate and $r'$ is the expected after-tax real interest rate.

Empirical tests for the Darby effect have proceeded on the basis of regression equations of the form

$$i = a_0 + a_1\overline{\Pi} + a_2 Z + e, \tag{4}$$

where $\overline{\Pi}$ is some measure of expected inflation based on past inflation and/or survey data, and $Z$ is a vector of standardizing variables ostensibly related to $i$. The test of the Darby hypothesis is then typically taken to be whether or not $\hat{a}_1$ exceeds one, since under the hypothesis $a_1$ equals $1/(1-T)$. Plausible values of $a_1$ suggested in the literature range in the neighborhood of 1.47 to 2.0, corresponding to marginal tax rates of 32–50 percent, respectively (see Tanzi, 1980, Feldstein, and Darby). Despite a large number of attempts to do so, relatively few values of $\hat{a}_1$ greater than one have been observed.[2] More often than not, estimates of $a_1$ are close to or less than one.

Response to these findings has been mixed. Several authors, such as Tanzi (1980), have accepted the statistical results more or less at face value, and declared that lenders have failed to foresee the effects of taxes on actual yields and thus have suffered from "fiscal illusion." Other observers, unable to accept the degree of ignorance or irrationality which such a conclusion implies, have sought statistical and/or theoretical explanations to reconcile these results with the hypothesis.[3]

---

*Department of Economics, California State University, Fullerton, CA 92634. This paper has benefited from conversations with Joyce Pickersgill and Lyle Kalish, and from the comments of a referee. Should it prove necessary, they are all exonerated from responsibility for the results.

[1] See Michael Darby (1975). The hypothesis has also been advanced independently by Martin Feldstein (1976) and Vito Tanzi (1976).

[2] Estimates of $a_1$ significantly in excess of one have been obtained by Jack Carr, James Pesando, and Lawrence Smith (1976), and Thomas Cargill and Robert Meyer (1980). Also, a referee has related that Cargill and Meyer have obtained estimates "...substantially larger than unity..." in a paper presented at the Federal Reserve Bank of San Francisco's November 1982 Academic Conference.

[3] Michael Melvin (1982) and Joe Peek (1982) have each derived a reduced-form Fisher equation the form of equation (4) from a macro model containing equation (2) as a structural equation. The idea in each paper was to allow inflation expectations to depress the real rate, thereby accounting for low values of $\hat{a}_1$. While both papers find evidence that taxes affect nominal interest rates, neither could be termed strong support for the Darby hypothesis. Melvin's estimated coefficients imply that nominal interest rates have (insignificantly) adjusted to a marginal tax rate of only 7.1 percent, while Peek's results are mixed, and unclear on the degree of adjustment.

## II. The Measurement Problem

Previous attempts to validate the Darby hypothesis have, I believe, failed to find conclusive support for the proposition due to the difficulty of observing $r'$ and $\Pi$. This difficulty arises because the relevant variables are not the after-tax real interest rate (which would be difficult enough) and the inflation rate, but the expected after-tax real interest rate and the expected inflation rate. This, of course, is recognized by those working in the area, and has been the motive force behind the increasingly sophisticated inflation expectation variables, and increasingly complex and lengthening vector of standardizing variables used in equation (4).

Rather than attempting to estimate the coefficient of expected inflation in equation (3), let us return to equation (2), and consider rewriting it as

$$i=(r'+\Pi)/(1-T). \tag{5}$$

This expression appears to differ only trivially from equation (3), but the difference is important. For although we cannot measure $r'$ or $\Pi$, we can measure $(r'+\Pi)$ as the nominal yield on tax exempt bonds.

Redefine $i$ above as $i_T$, the nominal yield on taxable bonds, and let $i_X$ be the nominal yield on tax exempt bonds of comparable maturity and risk. Then,

$$i_T=(r'+\Pi)/(1-T) \tag{5'}$$

under the Darby hypothesis, and

$$i_X=(r'+\Pi), \tag{6}$$

independent of the Darby hypothesis.

Note that it is not claimed in equation (6) that the Fisher equation is correct in the sense that actual tax exempt yields will rise in line with inflation, that is, that the after-tax real rate is constant and inflation expectations are correct. Rather, equation (6) says that, whatever the states of expectation concerning $r'$ and $\Pi$, they determine $i_X$. The expected values are, by definition, those that participants in the market act on.

Assuming the same state of expectations prevails among marginal purchasers of each type of security, we may substitute equation (6) into equation (5′) to obtain

$$i_T=i_X/(1-T). \tag{7}$$

The test of the Darby hypothesis is now whether or not the coefficient from the simple regression of $i_T$ on $i_X$ is significantly greater than one.

## III. Results

Equation (7) was estimated using quarterly averages for the years 1952–79, inclusive. The sample period was terminated in 1979 to avoid the credit controls and interest rate gyrations of 1980. The variable $i_T$ is the yield on one-year Treasury bills and $i_X$ the yield on one-year prime grade municipal bonds drawn from the *Federal Reserve Bulletin* and Salomon Brothers *Bond Market Roundup*, respectively.[4,5] Nominal yields on prime grade municipals ranged from a low of 0.63 percent, with a yield spread to the Treasury bills of 0.07 percentage points in 1954, to a high of 5.87 percent, with a yield spread to the Treasury bills of 5.29 percentage points in 1979. Equation (8) reports the estimated coefficients (with standard errors shown below in parentheses), with the variables expressed as percents:

$$\hat{i}_T=\underset{(.114)}{0.158}+\underset{(.038)}{1.63}\,i_X;\ R^2=.94. \tag{8}$$

This equation shows an unmistakable Darby effect.[6] On average over the sample period, a one percentage point rise in the nominal yield on one-year tax exempt bonds has been associated with a 1.63 percentage point rise in the nominal yield on one-year taxable

---

[4] Municipalities are not subject to taxes on their revenues, but presumably must pay at least the going market real after-tax yield, and need pay no more than this.

[5] The Salomon Brothers data was graciously supplied to me by Joyce Pickersgill.

[6] Note also that, independent of the tax effect, the estimated relation should pass through the origin, and it does.

bonds. This is sufficient to compensate lenders for an effective marginal tax rate on interest income of 38.7 percent.[7]

Recall, however, that $i_X$ measures only the sum $(r'+\Pi)$. Thus we can infer from these results neither the actual course of inflation expectations over the sample period, nor whether or not the expected real rate was depressed by expectations of inflation. Nonetheless, the presence of the Darby effect is clear.

It may be of some interest to know how the estimated relation fares over the sample period. Equations (9) and (10) report the results respectively of estimating equation (7) for the years 1952–65 and 1966–79 inclusive:

$$(9)\quad \hat{i}_T = \underset{(.118)}{-0.07} + \underset{(.069)}{1.77}\, i_X; \quad R^2 = .92,$$

$$(10)\quad \hat{i}_T = \underset{(.402)}{0.36} + \underset{(.104)}{1.57}\, i_X; \quad R^2 = .81.$$

The implied marginal tax rates are 43.5 percent for the first half of the sample period, and 36.3 percent for the latter half, each of which is similar to the estimate for the full sample period.

### IV. Summary and Conclusion

Previous studies attempting to observe the Darby effect have been plagued by the difficulty of measuring expected after-tax real interest rates and expected inflation rates. This problem is circumvented in the present study by using the nominal yield on tax exempt bonds to measure the sum of these two variables.

The growing yield spread between taxable and tax exempt bonds accompanying the rising nominal yield on tax exempt bonds is evidence of the Darby effect. Estimation of the relation between these yields for the 1952–79 period indicates that the nominal yield on taxable bonds has risen sufficiently to compensate lenders for an effective marginal tax rate in the neighborhood of 40 percent on interest income.

## REFERENCES

**Cargill, Thomas F.**, "Direct Evidence of the Darby Hypothesis for the United States," *Economic Inquiry*, January 1977, *15*, 132–34.

______ **and Meyer, Robert A.**, "The Term Structure of Inflationary Expectations and Market Efficiency," *Journal of Finance*, March 1980, *35*, 57–70.

**Carlson, John A.**, "Expected Inflation and Interest Rates," *Economic Inquiry*, October 1979, *17*, 597–608.

**Carr, Jack, Pesando, James E. and Smith, Lawrence B.**, "Tax Effects, Price Expectations and the Nominal Interest Rate," *Economic Inquiry*, June 1976, *14*, 259–69.

**Darby, Michael R.**, "The Financial and Tax Effects of Monetary Policy on Interest Rates," *Economic Inquiry*, June 1975, *13*, 266–76.

**Ezrati, Milton J.**, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates: Comment," *American Economic Review*, September 1982, *72*, 854–57.

**Feldstein, Martin**, "Inflation, Income Taxes, and the Rate of Interest: A Theoretical Analysis," *American Economic Review*, December 1976, *66*, 809–20.

**Fisher, Irving**, *The Theory of Interest*, New York: Macmillan, 1930.

**Lahiri, Kajal**, "Inflationary Expectations: Their Formation and Interest Rate Effects," *American Economic Review*, March 1976, *66*, 124–31.

**Melvin, Michael**, "Expected Inflation, Taxation, and Interest Rates: The Delusion of Fiscal Illusion," *American Economic Review*, September 1982, *72*, 841–45.

**Mundell, Robert A.**, "Inflation and Real Inter-

---

[7] If anything, this estimate seems likely to be biased downward. In addition to the errors-in-the-independent-variable problem, it turns out that municipal bonds are not quite tax exempt. Some purchasers of municipal bonds, residing in states other than the state of issue, may have been subject to state income taxes on the interest. In the presence of a Darby effect, we would now have $(1-T)i_T = (1-t)i_X$, $(0 \le t \le T)$; and in the regression of $i_T$ on $i_X$, we are estimating $(1-t)/(1-T)$. If $t$ is greater than zero and the regression coefficient taken as $1/(1-T)$, we have a downward-biased estimator of $T$.

est," *Journal of Political Economy*, June 1963, *71*, 280–83.

**Pearce, Douglas K.,** "Comparing Survey and Rational Measures of Inflation," *Journal of Money, Credit, and Banking*, November 1979, *11*, 447–56.

**Peek, Joe,** "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review*, December 1982, *72*, 980–91.

**Tanzi, Vito,** "Inflation, Indexation and Interest Income Taxation," Banca Nazionale del Lavoro *Quarterly Review*, March 1976, *29*, 54–76.

______, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates," *American Economic Review*, March 1980, *70*, 12–21.

______, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates: Reply," *American Economic Review*, September 1982, *72*, 860–63.

**Board of Governors of the Federal Reserve System,** *Federal Reserve Bulletin*, Vols. 38–65, Washington: 1953–80.

# The Impact of Minimum Wages on the Distributions of Earnings for Major Race-Sex Groups: A Dynamic Analysis

*By* JERE R. BEHRMAN, ROBIN C. SICKLES, AND PAUL TAUBMAN*

Since the passage of the Fair Labor Standards Act (FLSA) in 1937, the impact of minimum wage laws on the distribution of earnings has aroused considerable controversy. Advocates claim that a minimum wage aids in the attainment of equalitarian distribution goals by assuring a "fair" level of earnings to those workers at the low end of the wage distribution. Critics claim that the higher wage rate is more than offset by reduced employment among low-skilled workers, and no benefits are provided to disabled and other nonlabor market participants.[1] The social and economic importance of this debate has increased over time because of the secularly increasing coverage of the minimum wage laws. In 1956, only about one-half of the U.S. work force was in FLSA-covered jobs, but two decades later over four-fifths were covered.

Simple economic theory suggests that, with the possible exception of persons in monopolistic labor markets, effective minimum wage laws reduce the quantity demanded of laborers whose wage otherwise would be below the minimum wage.[2] This reduction in demand may be reflected in greater unemployment, in fewer hours worked by those still employed, and in lower labor force participation rates. The impact on the earnings of what Peter Linneman (1982) calls the "subminimum population" (i.e., those whose wage would be below the minimum level if there were no minimum wage) depends on the elasticities of demand and supply for labor in covered and noncovered occupations. If the demand for low-wage labor in the covered occupations is inelastic, earnings of the subminimum population in that sector increase with a rise in the effective minimum wage. However, even in this case the overall earnings of the subminimum population may decrease if they have a more than compensating fall in earnings from the uncovered sector which absorbs some of the displaced workers. Likewise, earnings of the subminimum population would fall if the demand for labor in the covered sector is elastic and the impact is not offset by a sufficiently elastic demand for low-wage labor in the uncovered sector. For the "superminimum" population (i.e., those with wages above the legal minimum), the effects depend upon the elasticities of substitution among various skill classes of laborers. For such reasons, the predictions of simple theory about the impact of minimum wages on earnings distributions are quite ambiguous.

A number of empirical studies have attempted to clarify the impact of minimum wages on unemployment and to a much lesser extent on earnings. Charles Brown, Curtis Gilroy, and Andrew Kohen (1982) and Kohen and Gilroy (1981) provide recent surveys. The available studies indicate some empirical regularities: teenage employment tends to fall by 1 to 3 percent for a 10 percent increase in the minimum wage and (in the words of Kohen and Gilroy), "the minimum wage has had small 'beneficial' effects on the distribution of income" (p. 25). However, even these regularities are not un-

---

*Department of Economics, 3718 Locust Walk CR, University of Pennsylvania, Philadelphia, PA 19104. This paper is an extension of our 1981a study. We thank the Minimum Wage Study Commission for research support, anonymous referees for useful comments, William Dreifke for dependable research assistance, and Brenda Liedtke, Monica Oberthaler, and Jennifer Goldenberg for bibliographical and secretarial contributions. As to responsibility for the analysis and interpretations herein, the usual caveat applies.

[1] This debate concentrates on earnings and does not place a value on leisure time.

[2] The impact on measured unemployment is ambiguous, however, if there is a response in labor force participation (see Jacob Mincer, 1976).

challenged.[3] Moreover, there are extensive ambiguities regarding such matters as the impact on various demographic groups (defined by age, sex, and race), the size (if any) of effects on unemployment, the role of schooling, the extent of longer-run effects, and the impact of changes in the minimum wage versus those of FLSA coverage.

Empirical investigations have not resolved many questions relating to the magnitudes and the timepaths of FLSA provisions in large part because of the following data inadequacies: 1) Time-series have been too short to allow much variation in real minimum wages and in coverage, to distinguish the effects of minimum wage variables from important macro variables like inflation, unemployment, and cohort size, or to allow estimation of adjustments over time. 2) Important micro controls for other variables, such as schooling and experience, have not been available. 3) Sample sizes have been too small to investigate differential effects among different sex-race demographic groups. 4) Variables to identify supply vs. demand factors have not been present.

Unfortunately, no data set currently exists which is free of all these inadequacies; however, the 1973 CPS-SSA-IRS Exact Match sample is a rich data source with some distinct advantages in comparison to previously used data sets. This sample contains annual earnings from Social Security Administration (SSA) records for the period 1951–76 for about 90,000 persons. We restrict our attention to those born between 1911 and 1936 who generally would not have retired by 1976 and generally were working in 1951 in order to investigate dynamic effects over the life cycle (see the Appendix). We exploit three major advantages of this data set.

First, the link between the CPS and the SSA records permits controlling for some important individual characteristics, such as schooling. This also permits us to explore whether the less schooled are affected differentially from the more schooled by changes in FLSA minimum wage levels or coverage, as is sometimes suggested (for example, see Feldstein and Ellwood).

Second, the large sample size permits subdivision into race and sex groups to see which groups, if any, benefit or lose from changes in minimum wage law provisions. We distinguish among white and black males and white and black females because of frequent hypotheses about the differential impact of FLSA provisions on these four demographic groups. The large sample size also permits direct estimation of the impact of changes in FLSA provisions on a number of characteristics of the earnings distributions, not just on the mean as usually is the case. We explore how minimum wage provisions alter four characteristics of the distributions of earnings: 1) the logarithm mean earnings relative to $GNP$ per laborer ($\ln \overline{Y}/(GNP/L)$); 2) the logarithm real cell mean earnings ($\ln \overline{Y}/P$); 3) within cell variance of the logarithm earnings ($\text{var} \ln Y$), and 4) the proportion below the poverty level for a two-adult, two-child family.

Third, the long time period permits observation of more changes in minimum wage rates and coverage, separation of cohort effects from time trends, control for macro conditions, and estimation of dynamic adjustments. All of these factors are important in obtaining better estimates of the impact of FLSA provisions on earnings distributions, but we focus particularly on the exploration of dynamic effects. Previous studies have concentrated on the effects on *current* earnings distributions of *concurrent* changes in FLSA provisions. But there are at least two good reasons to expect there to be important dynamic effects that exceed current year effects. First, time is required for many of the labor market adjustments noted in the second paragraph of this introduction. Second, if minimum wage laws affect employment at one point of time, they alter accumulated experience and the acquisition of tenure, on-the-job training and basic work habits, which may have an impact on subsequent earnings long after the original FLSA-induced employment change.

---

[3] For example, Martin Feldstein and David Ellwood conclude that "unemployment is not a serious problem for the vast majority of teenage boys" (1982, p. 30), which implies that male teenage unemployment caused by minimum wages must not be a serious problem. Also, Linneman concludes that minimum wages have contributed to inequality.

The important strengths of this data set are partially offset by the lack of data to permit identification of underlying supply and demand relations, by the truncation of earnings, by the less than complete coverage of the SSA, and by an incapacity to divide earnings into wage and hours-worked components. We believe the most serious problem arises from truncation. We deal with this problem by computing the appropriate statistical moments (and poverty share) for various cells, and then estimating the full sample moments using standard formulae. While our analysis is based on the moments of this aggregated data, it is not based on highly aggregated material since generally each cell covers one or two years of age, one or two grades of school, and one race and one sex.

Despite such limitations, the analysis of these data permits a substantial increment in our knowledge about the dynamic impact of the minimum wage system on earnings distributions, and it is earnings with which advocates of the minimum wage are largely concerned. In this paper we analyze these data to shed light on answers to the following important questions. What are the short- and long-run effects of FLSA provisions on the four characteristics of earnings distributions noted above for various sociodemographic groups? Do changes in coverage have different effects than changes in the minimum wage? Do the FLSA provisions benefit blacks, females, and the low-schooled as advocates of a minimum wage claim? Are lagged responses to FLSA provisions significant? Do they imply results different in magnitude or different in timing or both in comparison to the static analysis?

We organize our presentation as follows. Sections I and II discuss the specifications we estimate, Section III presents our estimates, and Section IV gives concluding remarks. Further discussion of the data and our resolution of the truncation problems are presented in the Appendix.

## I. Specification of Estimated Equations

Our dependent variables are the four characteristics of the earnings distributions defined in the introduction for each of the four race-sex demographic groups.[4] We interact schooling with the minimum wage and FLSA coverage variables because of hypotheses about different labor market effects for different schooling levels (for example, Feldstein and Ellwood). Otherwise, we specify simple linear relations. This generally linear specification can be considered a local first-order Taylor-series approximation to more complicated formulations.[5] To incorporate the dynamic possibilities, we specify each of our equations as a fairly general function with an endogenous variable ($Y_t^j$) and $k$ exogenous variables ($X_{it}^j$, $i=1,\dots,k$).

$$(1)\quad \alpha(L)Y_t^j=\sum_{i=1}^{k}\eta_i(L)X_{it}^j+\theta(L)U_t^j;\qquad (j=1,\dots,G),$$

where $L$ is the lag operator, $G$ is the number of cross sections, $U_t^j$ is a white noise process with constant variance, and

$$\alpha(L)=1-\sum_{j=1}^{P}\alpha_jL^j;\ \theta(L)=1-\sum_{j=1}^{Q}\theta_jL^j;$$

$$\eta_i(L)=\eta_{i0}-\sum_{j=1}^{M_i}\eta_{ij}L^j.$$

For each $j$ the function in (1), referred to in the stochastic processes literature as a transfer function, relates the dependent variable to lagged values of itself, current and lagged values of the exogenous variables, and to random effects or innovations. Before discussing our treatment of the dynamic lag structure, we consider the three major groups of exogenous variables which we include in our empirical relations.[6]

---

[4] The impact of minimum wage laws depends on factors such as the existence of market power, labor market segmentation, elasticities of labor demand and supply, elasticities of substitution among skill classes, discrimination, cohort effects, differential macro effects across markets, differentials in the extent to which the distributions of hours worked are changed, and dynamic adjustment patterns. Since no available data set permits the estimation of such a rich specification, we do not undertake an extensive theoretical discussion here. Our 1981b study and that by Brown, Gilroy, and Kohen give more extensive discussions.

[5] Of course the subdivision into race-sex groups allows general nonlinear effects across groups.

1) First are the minimum wage provision variables. We include the level of the federal legal minimum wage rate (in $) relative to the average nonfarm private wage rate $(W^m)$ and the proportion of the work force covered by FLSA $(C)$ as separate linear variables. We enter these two variables separately because of theoretically possible differential effects and to avoid the canceling of variations between them in a multiplicative composite formulation (for example, Brown, Gilroy, and Kohen, 1981; and Robert Mare and Christopher Winship, 1980).[7] Furthermore, as noted above, we include multiplicative interaction terms of schooling with the minimum wage ratio and coverage variables in order to explore whether or not the impact of the FLSA variables depends on schooling levels.[8]

2) Second, we use schooling and work experience to control for skills.[9]

3) Third, we include additional variables to control for various macro factors. We represent secular movements by a simple time trend and cohort effects by cohort dichotomous variables and by the cohort shares in the population. We use employment and inflation rates to control for cyclical macro fluctuations.[10]

## II. Statistical Treatment of Lags

Estimation of the system given in equation (1) involves three main steps. In the first, we use instrumental variables to estimate consistently the polynomials $\alpha(L)$ and $\eta_i(L)$, $i=1,\dots,K$. All of the contemporaneous and lagged values of the exogenous variables are used as instruments. Limitations imposed by the length of the time-series (26 years) and by limited variation in some of the exogenous variables restrict our choices of parameterizations. After preliminary analysis, we settled on a fairly flexible dynamic structure for which $P=2$ and $M_i \leqq 2$. For the explanatory variables, we include some variables in both contemporaneous and lagged form (i.e., $M_i=2$), but others only in contemporaneous form (i.e., $M_i=1$) because they vary little over time. Examples of the former category include the inflation rate, the minimum wage, coverage, the unemployment rate, minimum wage × schooling, and coverage × schooling; examples of the latter are dichotomous cohort variables and accumulated potential experience. In the next step we analyze different time-series representations of the estimated innovations for each cross section. The autocorrelation functions dampened monotonically and we were unable to reject the null hypothesis of $AR(1)$

---

[6] We use the same sets of variables in our estimates of the determinants of all four of the dependent variables. In addition, for the logarithm of variance earnings dependent variable, we explored the relevance of the variance of some of the right-hand side variables (for example, schooling). Despite some a priori appeal for such an approach, our estimates did not indicate significantly nonzero effects. This may be due to measurement error since for many observations the number of individuals on which such variances are calculated is very small. In any case, we decided that further pursuit of this tack would not be fruitful with this data set.

[7] In addition, we consider a multiplicative term between the minimum wage and coverage variables, but do not find empirical support for its importance.

[8] We use schooling as our best available proxy for skills. Schooling may be a good proxy either if it serves as a signaling device as is claimed by some, or if schooling actually adds to human capital as is frequently assumed (see fn. 9). Our procedure also is equivalent to allowing interaction terms between the minimum wage provisions, and the race and sex demographic variables used to define the samples.

[9] Behrman et al., (1980) demonstrate that the standard approach to estimating the returns to schooling in the form of earnings may result in substantial biases because of the failure to control for unobserved ability and motivation which arise out of family environment and genetic endowments. If the purpose is to control for unobserved factors as in the present study, rather than to estimate the returns to schooling, however, the apparent association between schooling and unobserved ability and motivation may be a blessing rather than a curse. In addition to providing micro controls, the inclusion of schooling and experience provides some base for being confident of the reasonableness of our procedures because our estimates of their coefficients in the static mean earnings relations are similar to those obtained in other studies (see our 1981b study, Appendix B).

[10] Because of possible simultaneity with the variables representing macro conditions, we have explored the treatment of the macro variables as being simultaneously determined. Since it is our conclusion that such treatment does not change the implication of the estimates on which we focus in this paper, we do not worry further about simultaneity in this paper.

errors. Therefore, the system we estimate is

$$(2)\quad Y_t^j = \alpha_1 Y_{t-1}^j + \alpha_2 Y_{t-2}^j + \sum_{i=1}^{k} \eta_i(L) X_{it}^j + \varepsilon_t^j$$

$$\varepsilon_t^j = \rho_j \varepsilon_{t-1}^j + \xi_t^j; \qquad j = 1,\dots,G,$$

where $\xi_t^j$ is a white noise process with constant variance $\sigma^2$.

The third step involves estimation of (2). We use an Aitken-type estimator to generate consistent, asymptotically normal estimates. To be more explicit we can stack the $G$ cross sections and rewrite (2) as

$$(3)\qquad Y = Z\delta + \varepsilon \quad \text{where } Y = \begin{bmatrix} Y_1^1 \\ \vdots \\ Y_T^1 \\ \vdots \\ Y_T^G \end{bmatrix};$$

$$Z = \begin{bmatrix} Y_1^1 & Y_1^1 & \vdots & X_1^1 \\ \vdots & \vdots & \vdots & \vdots \\ Y_{T-1}^1 & Y_{T-2}^1 & \vdots & X_T^1 \\ \vdots & \vdots & \vdots & \vdots \\ Y_{T-1}^G & Y_{T-2}^G & \vdots & X_T^G \end{bmatrix}; \; \varepsilon = \begin{bmatrix} \varepsilon_1^1 \\ \vdots \\ \varepsilon_T^1 \\ \vdots \\ \varepsilon_T^G \end{bmatrix};$$

and where $Y$ and $\varepsilon$ are of dimension $G(T-2)$ and $Z$ is of dimension $G(T-2)$ by $(\Sigma M_{i,i=1}^{k} + 2)$. The covariance structure is $E(\varepsilon\varepsilon') = \mathrm{diag}(R^1, R^2, \dots, R^G) = \Omega$, where $R_{t,s}^i = \sigma^2(\rho^i)^{|t-s|}$. The limiting distribution of the generalized least squares estimator is given by

$$(4)$$

$$\sqrt{N}(\hat{\delta} - \delta) \to \mathrm{N}\left(\left(Z'(\tilde{\Omega})^{-1}Z\right)^{-1} Z'(\tilde{\Omega})^{-1}Y, \left(\frac{Z'(\tilde{\Omega})^{-1}Z}{N}\right)^{-1}\right),$$

where $\tilde{\Omega}$ is formed from consistent estimates of $\rho_i$ based on the consistent first-step residuals and $N = G(T-2)$.

These estimates are asymptotically efficient relative to the instrumental variable estimates used in the first step, but are not as asymptotically efficient as maximum likelihood estimates. Computational resource constraints prevented us from using a Full-Information Maximum Likelihood estimator.[11] The block diagonal specification of $\Omega$ is appropriate only if the cells are independent. Using the consistent residuals from the first step, we calculated the $G(G-1)/2$ pairwise contemporaneous correlations and found them to be quite small and unambiguously insignificant.

To summarize the effects of the FLSA changes within this dynamic framework, we refer to multipliers which are calculated in the standard fashion as follows. Our model for a specific earnings distribution characteristic can be expressed as

$$(5)\qquad Y_t^* = Y_{t-1}^*\pi_0 + W_t^*\pi_1 + V_t^*,$$

$$\text{where}\quad Y_t^* = (Y_t, Y_{t-1}, X_t),$$

$$Y_{t-1}^* = (Y_{t-1}, Y_{t-2}, X_{t-1}),$$

$$W_t^* = X_t, \quad V_t^* = (\varepsilon_t, 0, 0),$$

and $\pi_0, \pi_1$ contain the appropriate elements of $\delta$.

From (5) we calculate the $n$th period impact multiplier and the $n$th period cumulative multiplier. The former is given by $\pi_1\pi_0^n$ and the latter by $\Sigma_{k=0}^{n}\pi_1\pi_0^k$. With a stable system such as ours, the total multiplier is $\Sigma_{k=0}^{\infty}\pi_1\pi_0^k = \pi_1(I-\pi_0)^{-1}$.

## III. Empirical Estimates

Here we discuss only the estimated impact of the minimum wage provisions on the four characteristics of the earnings distributions for the four race-sex groups, since the other variables are included primarily as controls.[12] We present these estimates as follows: Table 1 gives the estimated short- and long-run effects of the minimum wage ratio and FLSA coverage variables on each of the four characteristics of the earnings distributions for each of the four race-sex samples.[13] Table 2 gives "crossover" schooling levels at which the impacts change from negative to positive (and which are discussed further below). Table 3 gives the ratios of the long-run to the short-run responses at mean schooling levels.[14]

---

[11] We have, however, checked our estimates against those generated by the residual adjusted two-step estimator of Michio Hatanaka (1974). This estimator was originally proposed for a single cross section and is asymptotically efficient. The residual adjusted two-step estimator differs from ours in that the lagged value of the consistent residuals estimated in the first step is included as an additional regressor. Our results do not change in any substantial way with use of the Hatanaka procedure.

We summarize the implications of these estimates with reference to eight of their major characteristics.

*Widespread impact of minimum wage laws*: The estimates imply significant effects at least at the 10 percent level of the minimum wage law provisions on various dimensions of the earnings distributions for the four race-sex groups in 14 out of 16 cases (Table 1). The two exceptions are the variance of the logarithm earnings for black males and the proportion of black females below the poverty level.

*Similarity of impact on first two dependent variables*: The first dependent variable is the logarithm of mean earnings relative to *GNP* per laborer, and the second is the logarithm of deflated mean earnings. These two need not move together because of movements in the nonearnings and depreciation shares of *GNP* and in the labor shares for labor other than the type being considered in a particular case. However, our estimated effect of the minimum wage law provisions on the two are quite similar.

*Least impact on proportion below poverty level*: The estimates suggest that an expansion in coverage reduces the proportion of males below the poverty level significantly, with the effect damped significantly for white males with more schooling. They also suggest that the proportion of white females below the poverty level is reduced significantly by an increase in the minimum wage, more so in this case for the more schooled. But despite this evidence of significant effects on the proportion below the poverty level, the overall density of significant estimates is only about one-half as high for this characteristic of the earnings distributions as for the other three.

*More frequent significance of FLSA coverage than of minimum wage variable*: During our sample period, the changing FLSA coverage had more widespread significant effects on dimensions of the earnings distributions than did changes in the minimum wage ratio. There are 16 significantly nonzero variables related to coverage as compared to 10 for the minimum wage ratio. Or, to summarize the same pattern in a slightly different way, coverage has significant coefficient estimates in 10 of the 16 regressions, while the minimum wage has significant coefficient estimates in 7 of 16 (Table 1). This contrasts somewhat with the Brown, Gilroy, and Kohen result that coverage has been less important than the minimum wage in regard to employment effects.

But two points merit emphasis. First, the broader impact of coverage than of the minimum wage ratio on earnings over the sample period implied by our estimates is less than in our earlier static results (1981a). This provides one example of how static estimates may be misleading in comparison to the a priori more attractive dynamic ones —in this case implying a relative overemphasis of the importance of coverage relative to minimum wage levels in the sample period.

Second, as noted in the introduction, in the sample period coverage expanded considerably. But since coverage is now quite

---

[12] The complete estimates and some discussion of the other coefficients are in our 1981b study. Behrman and Taubman (1983) also discuss some related estimates regarding the impact of the macro unemployment and inflation variables on various characteristics of earnings distributions for different demographic groups, and our 1981a study discusses the estimated minimum wage law effects with a static version of relation (1).

[13] By short run, we mean the sum of the estimated coefficients of the contemporaneous and one-period lagged variables (i.e., the sum of appropriate elements of $\delta$ in equation (3)). By long run, we mean the total multiplier impact as defined earlier.

[14] Plots of the long-run effects of sustained changes in the minimum wage and FLSA coverage on the four earnings distributions' characteristics are available from the authors on request.

Table 1—Short- and Long-Run Effects of Minimum Wage ($W_m$) and FLSA Coverage ($C$), and Interaction with Schooling ($S$) on Four Characteristics of Earnings Distributions[a]

| Dependent and Exogenous Variables | White Males | | Black Males | | White Females | | Black Females | |
|---|---|---|---|---|---|---|---|---|
| | Short Run | Long Run | Short Run | Long Run | Short Run | Long Run | Short Run | Long Run |
| $\ln(\bar{Y}/GNP/L)$ | | | | | | | | |
| $W_m$ | $-.0047$ | $-.012$ | $-.0002$ | $-.00024$ | $-.0094^c$ | $-.040^c$ | $-.015$ | $-.037$ |
| $C$ | $-.66^b$ | $-1.78^b$ | $-.28^c$ | $-.28^c$ | $.125$ | $.54$ | $.27^b$ | $.66^b$ |
| $W_m^*S$ | $.001^b$ | $.0027^b$ | $.00072$ | $.00072$ | $.0012^c$ | $.005^c$ | $.0013$ | $.0029$ |
| $C^*S$ | $-.027$ | $-.069$ | $-.01$ | $-.0082$ | $.006$ | $.026$ | $.01^b$ | $.025^b$ |
| $\ln(\bar{Y}/P)$ | | | | | | | | |
| $W_m$ | $-.0003$ | $-.00081$ | $.0049$ | $.0071$ | $-.0085^c$ | $-.043^c$ | $-.013$ | $-.032$ |
| $C$ | $-.5^b$ | $-1.62^b$ | $-.15^c$ | $-.22^c$ | $.04$ | $.20$ | $.33^b$ | $.83^b$ |
| $W_m^*S$ | $.00091^b$ | $.0031^b$ | $.00077$ | $.0011$ | $.0011^c$ | $.0054^c$ | $.0013$ | $.0032$ |
| $C^*S$ | $-.022$ | $-.073$ | $-.034^c$ | $-.005^c$ | $.009$ | $.045$ | $.02^b$ | $.048^b$ |
| Var (ln $Y$) | | | | | | | | |
| $W_m$ | $-.088$ | $-.12$ | $.088$ | $.058$ | $-.112^b$ | $-.14^b$ | $-.024^c$ | $-.041^c$ |
| $C$ | $-5.29^b$ | $-6.90^b$ | $-3.10$ | $-2.04$ | $-4.55^b$ | $-5.59^b$ | $-2.75$ | $-4.66$ |
| $W_m^*S$ | $.0074$ | $.0097$ | $-.016$ | $-.010$ | $.011^b$ | $.013^b$ | $.0021$ | $.0035$ |
| $C^*S$ | $.53^c$ | $.69^c$ | $-.20$ | $-.14$ | $.28^c$ | $.35^c$ | $.18$ | $.29$ |
| Proportion < Poverty Level | | | | | | | | |
| $W_m$ | $-.0009$ | $-.0045$ | $-.0030$ | $-.0078$ | $.0003$ | $.0010$ | $.0004$ | $.0015$ |
| $C$ | $-.03^b$ | $-.15^b$ | $-.25^b$ | $-.66^b$ | $-.14$ | $-.48$ | $-.16$ | $-.59$ |
| $W_m^*S$ | $.00007$ | $.00035$ | $.00011$ | $.00029$ | $-.00011^c$ | $-.00037^c$ | $-.00017$ | $-.00062$ |
| $C^*S$ | $.007^b$ | $.035^b$ | $.017$ | $.045$ | $.0029$ | $.0097$ | $-.017$ | $-.063$ |

[a]Short run refers to the sum of the coefficient estimates on the contemporaneous and one-period lagged terms. Long run refers to the total multiplier, as defined in Section III. The dependent variables are defined in Section I and the right-side variable in Section II. For more details see Behrman, Taubman and Sickles (1981b).

[b,c]Underlying coefficient significantly nonzero at the 5 percent (10 percent) level.

widespread, even if changes in coverage were more important than changes in the minimum wage in the sample period, the reverse may be the case in the future.

*Wider significant impact on whites than on blacks, but more favorable effects on low-schooled blacks*: A common perception seems to be that blacks are relatively more affected by the minimum wage laws. In terms of statistically significant coefficients we find otherwise, with about twice as many significant estimates for whites than blacks. Brown, Gilroy, and Kohen report a similar pattern for the employment and unemployment effects of minimum wage provisions.

For some purposes it is more appropriate to compare the absolute magnitudes of the coefficient estimates in Table 1. The distributions of earnings are compressed more at lower schooling levels for whites than for blacks.[15] For males with lower schooling levels, increased minimum wage provisions reduce the mean earnings (and the mean earnings relative to *GNP* per laborer) more for whites than for blacks. On the other hand, the estimated absolute impacts on these characteristics of the distributions for females tend to be larger for blacks than for whites,[16] particularly for the more important positive coverage effects. Furthermore, blacks tend to be affected more than whites with regard to the proportion below the poverty level, with the larger reduction of those with low schooling below this cutoff when coverage is increased most striking. Thus, though there are more significant effects for whites than for blacks, in terms of the magnitude of the estimated coefficients there seems to be approximate balance of the effects on the two races.

---

[15] The effects at lower schooling levels are indicated by the coefficient estimates of the minimum wage or of the coverage variables alone, since the interaction terms with schooling become very small if schooling is very low. As discussed below, sometimes the directions of these effects are reversed at higher schooling levels.

[16] The only exception to this statement is for the short-run effect of the minimum wage level on female average earnings relative to *GNP* per laborer.

In addition to the question of whether blacks or whites are affected more by the minimum wage law provisions is the question whether one race is affected more favorably than is the other. This has been a matter of some controversy. Linda Dachter and Glenn Loury (1981), for example, conclude that minimum wage laws favor white over black families. Linneman, in contrast, concludes that there is "surprisingly little influence of race...on expected earnings" (p. 468) in his study of minimum wage effects. Our estimates suggest that blacks with lower schooling levels tend to be affected more favorably than are whites with lower schooling levels. For mean earnings relative to *GNP* per laborer, real earnings and proportion below the poverty level variables, the estimated impacts tend to be more positive (or less negative) for blacks.

*About the same impact on males and females*: Another common perception is that females are more affected by minimum wages than are males. Once again there is some controversy about whether females benefit relatively (for example, Edward Gramlich, 1976 and Donald Parsons, 1980) or lose (Linneman) from the impact of minimum wages. We find very weak evidence of more widespread effects for females than for males (14 vs. 12 significant coefficients). But again the relative absolute sizes of the impacts are mixed, with about as many that are larger for females as for males. Our estimates also imply that neither sex is favored strongly by increases in the minimum wage, but that females have gained relatively by expansion of FLSA coverage.

*Importance of interactions with schooling, with negative implications for the less schooled*: The minimum wage rates affect most directly those with low wages. Minimum wages also may induce a substitution towards high-skilled labor. To account for this we have interacted our minimum wage variables with years of schooling. Our dynamic results contain almost as many significant coefficients on the interaction terms with schooling as on the linear minimum wage and coverage variables (12 vs. 14).[17] In all but two of the cases in which the interaction terms have significant coefficient estimates in Table 1, their signs imply that raising the relative minimum wage or the extent of coverage hurts those with little schooling in comparison to those with more schooling.[18] Most previous studies do not highlight the schooling and minimum wage law interactions, so we cannot directly compare this result with those from previous studies. However, it does contrast with the spirit of the Kohen and Gilroy survey summary noted in the introduction regarding the "beneficial" distribution effects.

An interesting feature of most of our coefficient estimates of the interaction terms is that they are opposite in sign to those of the linear minimum wage rate or coverage terms. This implies a crossover schooling level, at which the impact of the minimum wage provision on the relevant characteristics of the earnings distribution changes sign. Table 2 gives the crossover schooling levels implied by the estimates in Table 1, as well as the mean schooling levels for each sample. In some cases, the crossover levels are sufficiently low or high, so that the sign of the estimated total impact is the same for virtually everyone in the sample; though, of course, the estimated magnitude differs depending on the schooling level. But, in many cases, the crossover schooling levels are such that substantial proportions of the samples are both above and below them, implying different signs of effects for those with low schooling levels versus those with high

---

[17] In our static model (1981a), the schooling interaction terms have significant coefficient estimates even more frequently. Thus, in the present study, the multicollinearity introduced by the dynamic terms may be disguising some additional significant interaction effects.

[18] The two exceptions are the coverage-schooling interactions in the relation for the proportion of white males below the poverty level and in the relation for the logarithm of deflated mean earnings for black males.

TABLE 2—MEAN SCHOOLING LEVELS AND CROSSOVER SCHOOLING LEVELS AT WHICH TOTAL RESPONSES CHANGE SIGN[a]

| | | White Males ($\bar{S}=11.7$) | | Black Males ($\bar{S}=9.4$) | | White Females ($\bar{S}=11.5$) | | Black Females ($\bar{S}=10.0$) | |
|---|---|---|---|---|---|---|---|---|---|
| Dependent Variable | Minimum Wage Change | Short Run | Long Run | Short Run | Long Run | Short Run | Long Run | Short Run | Long Run |
| $\ln(\bar{Y}/GNP/L)$ | $W_m$ | 4.7 | 4.4 | .3 | .3 | 7.8 | 8.0 | 11.6 | 12.8 |
| | $C$ | – | – | – | – | – | – | – | – |
| $\ln(\bar{Y}/P)$ | $W_m$ | .3 | .3 | – | – | 7.7 | 8.0 | 10.0 | 10.0 |
| | $C$ | – | – | – | – | – | – | – | – |
| $\mathrm{Var}(\ln Y)$ | $W_m$ | 11.9 | 12.4 | 5.5 | 5.8 | 10.2 | 10.8 | 11.4 | 11.7 |
| | $C$ | 10.0 | 10.0 | – | – | 16.3 | 16.0 | 15.3 | 16.1 |
| Proportion < | $W_m$ | 12.9 | 12.9 | 27.3 | 26.9 | 2.7 | 2.7 | 2.4 | 2.4 |
| Poverty Level | $C$ | 4.3 | 4.3 | 14.7 | 14.7 | 48.3 | 49.5 | – | – |

[a] Calculated from estimates in Table 1 in cases in which the sign of the interaction term coefficient estimate differs from that of the linear term. $\bar{S}$ refers to the sample mean schooling.

TABLE 3—ESTIMATED RATIOS OF LONG-RUN TO SHORT-RUN RESPONSES TO CHANGES IN MINIMUM WAGE PROVISIONS FOR FOUR CHARACTERISTICS OF EARNINGS DISTRIBUTIONS[a]

| Dependent Variable | White Males | Black Males | White Females | Black Females |
|---|---|---|---|---|
| $\ln(\bar{Y}/GNP/L)$ | 2.7 | 1.0 | 4.3 | 2.4 |
| $\ln(\bar{Y}/P)$ | 3.2 | 1.4 | 5.0 | 2.5 |
| $\mathrm{Var}(\ln Y)$ | 1.5 | 0.7 | 1.2 | 1.7 |
| Proportion < Poverty Level | 5.0 | 2.6 | 3.3 | 3.7 |

[a] See Section II for definitions of long-run total multipliers. These ratios are calculated at mean schooling levels.

schooling levels. For example, for all but black males, increases in the minimum wage rate tend to compress the distribution of earnings for those with low schooling, but tend to expand the variance for those with higher schooling levels—and the crossover schooling levels are close to the sample means so substantial numbers are both above and below the critical schooling levels.

*Widespread and substantial differences between short- and long-run responses*: Table 3 summarizes the estimated ratios of long- to short-run responses at the sample mean schooling levels. In some cases these ratios are relatively small. Noteworthy in this respect are those for black males (except for the proportion below the poverty level) and for the variance of the logarithm earnings. But frequently the ratios indicate that the long-run responses are substantially greater than are the short-run responses. Therefore, consideration of the short-run responses alone, as in most analyses of the impact of minimum wage laws, may understate the effects.[19] Dynamics apparently are important because of adjustment lags and long-run effects of short-term phenomena. As a result, in the majority of cases the estimated long-run effects are at least twice the magnitude of the short-run effects, and in some cases as much as five times as high!

---

[19] Or, if the current static variables represent the omitted lagged dynamic ones, the static analysis may understate the adjustment periods.

There are also some interesting patterns in the distributions of sustained long-term effects of minimum wage and FLSA coverage on our measures of earnings distributions for the mean schooling levels.

The long-term effects of sustained minimum wage changes are reached in about fifteen years, although the time horizon is markedly shorter for black males and black females and for the percentage below the poverty level. The demographic groups for whom the long-term income effects are the most positive are white males and females. However, the minimum wage does seem to have reduced the percentage of black males and females with mean schooling below the poverty level markedly with very little effect on white males. White males and females, nevertheless, seem to have benefited the most with respect to their level of per capita earnings relative to black males and females. It is also noteworthy that the labor share of *GNP* has been increased by higher minimum wage rates since the slight decline for black females is much more than offset by the larger increases for the other three groups.

The long-term effects of coverage have somewhat different patterns than those of the minimum wage. As with the minimum wage, the long-run effects are fully attained in fifteen years. However, extended coverage has increased the mean earnings going to both black and white females. Moreover, females seem to have benefited substantially from extended coverage in reducing their percentage below the poverty level. Furthermore, coverage seems to have reduced the dispersion in the income distribution. The var($\ln Y$) is substantially reduced for all but white males, whose variance increased marginally. Finally, coverage has had a less positive impact on the labor share in *GNP* than has the minimum wage rate since the gains for females are largely offset by the losses for white males.

## IV. Conclusions

Are dynamics important? Our estimates must be qualified because they are based on estimated cell moments from truncated earnings distributions without identification of the underlying structural relations. Subject to this caveat, our results indicate that dynamics are very important in understanding the impact of minimum wage law provisions on earnings distributions. In the majority of the cases, the long-run effects are at least twice the magnitude of the short-run effects and, in many cases, the adjustment periods are quite long. The empirical importance of dynamics suggests that time is required for important labor market adjustments, and short-run employment and training impacts have longer-run persistent effects. Therefore, previous studies with their focus on concurrent "static" effects tend to underestimate both the magnitudes and the lags of impacts. Moreover, these biases differ among race-sex groups and among earnings distribution characteristics, so static analysis may not even provide a good indication of the relative effects of FLSA policies across demographic groups or across earnings distributions characteristics.

What do the preferred dynamic estimates indicate about the pattern of the impact of FLSA provisions? There is substantial support for the need to distinguish between changes in the real minimum wage and changes in FLSA coverage, as Mare and Winship and Brown, Gilroy, and Kohen argue. During the sample period, the effects of these two differed significantly, with expanded coverage having somewhat broader impact. But, in the future, extension of coverage will be less important since most of the work force currently is covered.

What are the effects of the minimum wage law provisions on earnings distributions[20] for groups who are the supposed beneficiaries in the arguments of many advocates—blacks, females, the low schooled? The answer to this question is mixed. There are some positive effects. Expansion of coverage has had a strong positive effect for females, a result which reinforces that of Gramlich, but contrasts sharply with the conclusion of Linneman on the bases of static estimates which focus on the minimum wage ratio alone. Increases in the minimum wage ratio have reduced the proportions of females and blacks below the poverty level, and have increased the real mean earnings for white females and black males.

---

[20] Throughout this paper we consider only the distributions of earnings among those employed in SSA-covered occupations. The impact on the earnings of all labor force participants may be different, and the impact on distributions for everyone may be still different yet.

But there also are some important negative dimensions to the answer to this question. The expansion of coverage has not benefited black males. An increase in the minimum wage ratio has not increased the real mean earnings for black females, but has done so for white males. Most important, generally the changes in the FLSA provisions have benefited the more schooled relative to the less schooled, a finding which reinforces Linneman's static conclusions in this respect, but contrasts with the Kohen and Gilroy survey conclusion that there are "beneficial" distribution effects.

Thus, at best, a summary evaluation of the impact of federal minimum wage law on the earnings of the working poor, the supposed beneficiaries, is mixed. In addition, there are other probable negative effects pertaining to employment and efficiency outside of the scope of this study.[21] Moreover it has affected differentially individuals with identical characteristics (except for demographic features like race and sex, which legally—and in our judgment, morally—should be irrelevant). Thus, we agree with Kohen and Gilroy, and with Linneman that federal minimum wage policy on net probably has not been good policy.

## APPENDIX: DATA

Many of our variables are from standard sources: minimum wage rates and FLSA coverage are from the Employment Standards Administration; unemployment and population are from Current Population Reports Series P-25 and Bureau of Labor Statistics Bulletin No. 2000, respectively; and inflation, *GNP*, and its components come from the *Survey of Current Business*.

Also we utilize the rich new data source mentioned in the text: the 1973 CPS-SSA-IRS Exact Match sample which provides a link between individual characteristics (age, race, sex, schooling) from the 1973 CPS and SSA annual earnings records for 1951–76. We use these data to obtain micro characteristics of individuals and to construct our four dependent variables relating to the race-sex earnings distributions. To construct the dependent variables, we first aggregate the 90,000 individuals into age, race, sex, and schooling cells of at least 150 individuals each, and then calculate the necessary means, variances, and proportions below the poverty line for all such cells. The aggregation into cells has the advantages of enabling direct representation of all of the characteristics of the distributions of interest, of controlling for schooling and experience within a reduced-form framework (see Section I), and of reducing a vast quantity of data (i.e., an average of about 15 years of earnings for each of the 90,000 individuals) to a more manageable size. The cost is that some interesting details, such as those emphasized by Feldstein and Ellwood for male teenagers, may be lost. We consider only those cohorts for each race-sex group whose year of birth was between 1911–34 which yields 26 years and up to 63 cross sections. Because of our desire to impose the same autoregressive parameterization on the error term, we exclude those whose year of birth was before 1911 or after 1934, since their members were not normally in the labor force in all years of the 26-year period.

There are two major limitations of these data. First, they do not permit the identification of the underlying structural labor market supply and demand relations. We do not have separate data on wages and hours, and we are not able to distinguish working in non-SSA-covered occupations from being unemployed or not being in the labor force.[22] We finesse the last problem by assuming that SSA-covered employment is a random draw[23] from the overall earnings distribution and use the samples' four moments.

---

[21] For example see Mincer and Linda Leighton (1980), Michael Wachter and Jeffrey Perloff (1980), and Richard Zeckhauser (1971).

[22] For an analysis of wage and hours effects in union and nonunion sectors vis-à-vis covered and noncovered sectors, see Perloff and Sickles (1983).

[23] Since 1956 about 90 percent of the labor force has been covered by the SSA. The groups not covered by the

Second, prior to 1978, the SSA only recorded a person's earnings (from each employer) up to the taxable ceiling. The earnings data therefore are drawn from a truncated distribution. It is well known that truncation may cause the estimated moments to be biased. A standard way to handle truncated distributions is to assume that the sample is drawn from a particular theoretical distribution in which the relationship between truncated and nontruncated moments are a function of known information. We follow this procedure and use the lognormal distribution, which several economic studies have concluded provides a tolerable fit to the earnings distribution, to estimate the full distribution moments for each cell.[24] For further details concerning these data see our 1981b study.

## REFERENCES

**Anderson, Theodore W. and Hsiao, Cheng,** "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, September 1981, *76*, 598–606.

**Behrman, Jere R. and Taubman, Paul,** "Redistribution of Earnings by Unemployment and Inflation," in F. G. Adams and B. Hickman, eds., *Global Econometrics: Essays in Honor of Lawrence R. Klein*, Cambridge: MIT Press, 1983.

**______, ______, and Sickles, Robin C.,** (1981a) "The Impact of Minimum Wages on the Distributions of Earnings for Major Age-Race-Sex Groups," mimeo., Philadelphia: University of Pennsylvania, 1981.

**______, ______, and ______,** (1981b) "The Short- and Long-Run Effects of Minimum Wages on the Distribution of Income," *Report of the Minimum Wage Study Commission*, Vol. VII: *Effects of the Minimum Wage on the Distribution of Income*, Washington: USGPO, 1981, 63–112.

**Behrman, Jere R. et al.,** *Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment, and Schooling*, Amsterdam: North-Holland, 1980.

**Brown, Charles C., Gilroy, Curtis L. and Kohen, Andrew I.,** "Time-Series Evidence on the Effect of the Minimum Wage on Teenage Employment and Unemployment," *Report of the Minimum Wage Study Commission*, Vol. V: *Effects of the Minimum Wage on Employment and Unemployment*, Washington: USGPO, 1981.

**______, ______, and ______,** "The Effects of the Minimum Wage on Employment and Unemployment," *Journal of Economic Literature*, June 1982, *20*, 487–528.

**Datcher, Linda P. and Loury, Glenn C.,** "The Effect of Minimum Wage Legislation on the Distribution of Family Earnings Among Blacks and Whites," *Report of the Minimum Wage Study Commission*, Vol. VII: *Effects of the Minimum Wage on the Distribution of Income*, Washington: USGPO, 1981, 125–46.

**Feldstein, Martin and Ellwood, David T.,** "Teenage Unemployment: What is the Problem?," in Richard B. Freeman and David A. Wise, eds., *The Youth Labor Market Problem: Its Nature, Cause and Consequences*, Chicago: University of Chicago Press, 1982, 17–34.

**Gramlich, Edward M.,** "Impact of Minimum Wages on Other Wages, Employment and Family Incomes," *Brookings Papers on Economic Activity*, 2:1976, 409–61.

**Hatanaka, Michio,** "An Efficient Two-Step Estimator for the Dynamic Adjustment Model with Autoregressive Errors," *Jour-*

---

SSA range over the income distribution since the U.S. government is the single largest exempt employer. Thus we doubt that major biases are introduced by this assumption. Also see Fritz Scheuren and H. Lock Oh (1975).

[24] If the distribution is lognormal, the relationships between the truncated and full sample are given by truncated mean $= \mu - \sigma r((c-\mu)/\sigma)$ and truncated variance $= \sigma^2\{-1 - r((c-\mu)/\sigma)[((c-\mu)/\sigma) + r((c-\mu)/\sigma)]\}$ where $r((c-\mu)/\sigma) = f((c-\mu)/\sigma)/F((c-\mu)/\sigma)$, $f$, and $F$ are the density and cumulative functions of the normal distribution, and $\mu$ and $\sigma^2$ are the full sample mean and variance. This system is easy to solve for the full sample mean and variance, since the term in brackets in the equation for truncated variance depends on terms which are transforms of the fraction of people below the ceiling. Given the solution of the latter equation for $\sigma^2$, it is easy to solve the former equation for an estimate of the full mean.

*nal of Econometrics*, September 1974, *2*, 199–220.

**Kohen, Andrew I. and Gilroy, Curtis L.,** "The Minimum Wage, Income Distribution, and Poverty," *Report of the Minimum Wage Study Commission*, Vol. VII: *Effects of the Minimum Wage on the Distribution of Income*, Washington: USGPO, 1981, 1–30.

**Linneman, Peter,** "The Economic Impacts of Minimum Wage Laws: A New Look at an Old Question," *Journal of Political Economy*, June 1982, *90*, 443–69.

**Makridakis, S. and Wheelwright, S. C.,** *Forecasting Methods and Applications*, New York: John Wiley & Sons, 1978.

**Mare, Robert D. and Winship, Christopher,** "Changes in the Relative Labor Force Status of Black and White Youths: A Review of the Literature," Working Paper No. SR26, University of Wisconsin-Madison, 1980.

**Mincer, Jacob,** "Unemployment Effects of Minimum Wages," *Journal of Political Economy*, July 1976, *84*, S87–S105.

______ **and Leighton, Linda,** "The Effects of Minimum Wages on Human Capital Formation," Working Paper No. 441, National Bureau of Economic Research, Cambridge, 1980.

**Palm, Franz and Zellner, Arnold,** "Large Sample Estimation and Testing Procedures for Dynamic Equation Systems," *Journal of Econometrics*, December 1980, *12*, 251–83.

**Parsons, Donald O.,** *Poverty and the Minimum Wage*, Washington: The American Enterprise Institute, 1980.

**Perloff, Jeffrey and Sickles, Robin C.,** "FIML Estimation of Union Wage and Hours Differentials in the Construction Industry: A Nonlinear Limited Dependent Variable Approach," mimeo., University of Pennsylvania, 1983.

**Scheuren, Fritz and Oh, H. Lock,** "Fiddling Around with Nonmatches and Mismatches," *Proceedings* of the Social Statistics Section of the American Statistical Association, 1975.

**Wachter, Michael and Perloff, Jeffrey,** "The Productivity Slowdown: A Labor Problem?," Boston Federal Reserve Bank Conference Series, June 1980.

**Zeckhauser, Richard J.,** "Optimal Mechanisms for Income Transfer," *American Economic Review*, June 1971, *61*, 324–34.

# The Price-Adjustment Process for Rental Housing and the Natural Vacancy Rate

*By* KENNETH T. ROSEN AND LAWRENCE B. SMITH*

Traditional economic and housing market analyses have ascribed a close connection between excess demand, as reflected in the deviation of the actual vacancy rate from some long-run normal or optimal vacancy rate, and changes in the price of rental housing services. This relationship was clearly described by David Blank and Louis Winnick (1953), and reiterated by Sherman Maisel (1963) and Ray Fair (1972). However, neither this relationship nor the determinants of the optimal or natural vacancy rate have been empirically demonstrated for the United States.[1] Moreover, studies by Frank de Leeuw and Nkanta Ekanem (1971), Arthur Eubank and C. F. Sirmans (1979) and Ira Lowry (1981) question the validity of this price-adjustment mechanism.

In contrast to these studies, this paper confirms the view that rental price changes are significantly affected by excess supply or demand in the rental market. We empirically investigate this relationship separately for seventeen U.S. cities, by a pooled time-series, cross-section analysis. We demonstrate that a close relationship exists between variations in rental prices and deviations of actual vacancy rates from normal vacancy rates. We also develop and estimate a cross-section model to explain natural vacancy rates, using those vacancy rates generated by the price-adjustment model.

The paper proceeds as follows. In Section I, we develop and show estimates of a model of rental price adjustment for the seventeen cities, and demonstrate the crucial role played by vacancies in the price-adjustment process. In Section II, our empirical results are used to generate long-run natural or optimal vacancy rates for fourteen of the cities, and a model is developed to explain these results.

## I. The Rental Price-Adjustment Process

The price-adjustment mechanism and the rental housing market can be viewed as operating in a typical stock-flow manner.[2] At any one time there is a stock of rental housing units providing housing services and a demand for these services. If we assume, as is usual,[3] that a standardized unit of housing stock yields a unit of housing services during each period of time, then the rent is the price of the flow of services from one standard dwelling unit, and the demand and supply of housing services can be considered as the demand and supply of units of housing stock. Although the size of the standardized rental housing stock in any period is increased by newly completed or converted rental dwellings and diminished by removals, demolitions and depreciation, the annual change in the stock is relatively small,[4] and hence the stock may be considered as fixed in the short run.

The demand for rental stock is usually assumed to depend upon a variety of variables including demographic variables (such as the number of families, tendency of the population to form nonfamily households,

---

*Professor of Economic Analysis and Policy, School of Business Administration, 350 Barrows Hall, University of California, Berkeley, CA 94720, and Professor, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Canada M5S 1A1. This research was conducted at the Center for Real Estate and Urban Economics, University of California-Berkeley, and was funded by the Center and by the Social Sciences and Humanities Research Council of Canada, Research Award 451-81-2999 to Smith.

[1] Smith (1974) demonstrated a similar relationship between the percentage change in rents and the vacancy rate for Canada.

[2] Much of the next three paragraphs is based on Fair and Smith.

[3] See, for example, Richard Muth (1960) and Edgar Olsen (1969).

[4] Approximately 1.5–2.0 percent a year.

and age composition of the population),[5] permanent real disposable income, the price of rental accommodation, the user cost of owner-occupied housing, the price of alternative goods and services, the cost and availability of mortgage credit, and consumer preferences.[6] These demand and supply functions interact to determine the level of rents and the stock of vacant rental units.

However, since numerous frictions and imperfections cause the market to adjust slowly, the rent level determined through this process may not completely clear the market in the sense that actual vacancies equal the normal or optimal vacancies. The natural or optimal vacancy rate, analogous to the natural unemployment rate, is defined by market factors such as the cost of holding inventory, search costs, the variability of demand, and the costs of recontracting. Market frictions such as high transactions and search costs, slow supply responses, credit market imperfections and the existence of long-term contracts may all impede the quick adjustment of rents.[7] If rents are such that the housing stock demanded exceeds the available supply less the normal level of vacancies, then vacancies will be less than normal and upward pressure will be exerted on rents. This will bring forth new construction and the conversion of existing units, as well as reduce demand from existing renters. Analogously, if rents are such that the housing stock demanded is less than the available supply less the normal level of vacancies, vacancies will be larger than normal, downward pressure will be exerted on rents, and new construction will be lower than in the market-clearing case. The speed at which the market moves toward equilibrium depends, among other things, upon the supply-side response and speed-of-rental price adjustment. This discussion implies that the rate of change of rents depends upon the vacancy rate, and that variations in the arguments in the demand function or the supply will be reflected initially in the vacancy rate (although they may also exert some direct effect on the rate of change in rents over the long term).

## A. *The Model*

This discussion can be summarized in an empirically testable model as follows. The demand for rental housing services and hence for the rental housing stock, $D$, may be assumed to be a function of the rent per unit of housing services, $R$; the user cost of homeownership, $U$; real income per household, $Y$; the price level, $P$; and demographic variables, $Z$, as set out in

$$D=d(R,U,Y,P,Z). \tag{1}$$

Since the supply of rental units, $S$, is assumed to be fixed in the short run, the level of vacancies, $VL$, is the difference between the demand and supply of rental units, as in

$$VL=S-D, \tag{2}$$

and the vacancy rate, $V$, is the ratio of the number of vacant units to the supply of rental units as shown in

$$V=VL/S=1-(1/S)d(R,U,Y,P,Z). \tag{3}$$

The rental price adjustment mechanism is set out in equation (4). In this mechanism, the excess demand or supply for rental housing, defined as deviations in the actual vacancy rate from the natural vacancy rate, $V^n$, determines variations in the real rent within some critical zone of occupancy.[8] The rate of change of nominal rents, $\dot{R}$, is thus a function of the excess demand or supply for rental housing and the rate of change of total operating expenses, $\dot{E}$, where $\dot{E}$ reflects nominal price influences on $\dot{R}$:

$$\dot{R}=r(\dot{E},V^n-V). \tag{4}$$

---

[5] For a discussion of the impact of demographic variables, see Robert A. Gordon (1956) and Dwight Jaffee and Rosen (1979).

[6] For a discussion of rental demand and tenure choice, see Douglas Diamond (1978); Patric Hendershott and James Schilling (1981); and Harvey Rosen and Rosen (1980).

[7] For a discussion of some of these frictions, see de Leeuw-Ekanem, Fair, and Maisel.

[8] See Blank and Winnick. The specification holds only within the normal or critical zone of vacancies since the model collapses as vacancies approach zero.

If we assume the natural vacancy rate is constant over our estimation period for any given city, $V^n$ may be incorporated into the intercept, and the estimating model may be written as

$$\dot{R} = b_0 + b_1\dot{E} - b_2V. \tag{5}$$

Since the market-adjustment lag depends upon the supply response and other institutional arrangements which vary between cities,[9] a variable lag structure was used between cities when estimating equation (5).

### B. *The Estimated Results*

Equation (5) was estimated for each of the seventeen cities[10] for which a sufficiently large data sample was available over the period from 1969 to 1980. The results are presented in Table 1.

The data on vacancy rates and expenses were derived from the Institute of Real Estate Management, *Annual Income/Expense Analysis for Apartments*.[11] Operating expenses were calculated in two stages. First, unweighted averages of the rent per square foot and of the operating expenses per square foot for four apartment building types (elevator buildings, low rise buildings with 12–24 units, low rise buildings with 25+ units, and garden-type buildings) were calculated annually for each city. Second, the ratio of these operating expenses to rents was then multiplied by the Bureau of Labor Statistics index of apartment rents for each city to generate an index of apartment building operating costs, $E$, for each city. The vacancy variable $V$ for each city was calculated as the annual unweighted average of the percentage of gross potential income not collected for the four apartment building types in the *Annual Income/Expense Analysis for Apartments*. Finally, the rental variable $R$ was taken directly from the Bureau of Labor Statistics index of apartment rents for each city.

The results of our estimation indicate that vacancies were significant in explaining the percentage change in rents at the 95 percent level in thirteen cities (in two additional cities, the $t$-statistics were over one), and were insignificant in only two cities. Given the relatively small size of the sample for each city and the variance in the buildings included in the survey over time, these results are relatively impressive. To further test the validity of our specifications, the city data were pooled and a combined cross-section, time-series regression was estimated with city dummy variables, $D_1 \cdots D_{16}$, introduced for all but one of the cities. This result, excluding the city dummy variables, is shown in the last line of Table 1, and strongly supports the appropriateness of our model. Consequently, our results strongly confirm the traditional view that vacancy rates within some critical zone of occupancy are pivotal in explaining the price of housing services.

## II. The Natural or Optimal Vacancy Rate

It has been hypothesized by de Leeuw and Ekanem (1971, p. 812) that their failure to obtain a significant relationship between rents and vacancy rates across metropolitan areas could be due to the fact that the variation in vacancy rates among metropolitan areas reflects differences in the "normal" vacancy rates rather than different degrees of market tightness. To examine the possibility of intercity variations in the natural vacancy rate, we used the empirical results in Table 1 to calculate a natural vacancy rate for each city and developed a model to explain these rates.

In a manner analogous to the labor market, the housing market requires some normal stock of vacant units to facilitate the search processes of buyers and sellers in the market.[12] These processes and the "unpre-

---

[9] These differences arise from a variety of factors including differing zoning regulations, planning procedures, and other land use controls.

[10] Baltimore, Boston, Chicago, Cincinnati, Cleveland, Dallas, Denver, Houston, Kansas, Los Angeles, Milwaukee, Minneapolis, New York, Philadelphia, Pittsburgh, St. Louis, and Washington.

[11] Data for 1974 were estimated by interpolation since the data were not available for that year.

[12] See, for example, Armen Alchian (1970) for a discussion of this process in the labor market.

TABLE 1—DETERMINANTS OF THE PERCENTAGE CHANGE IN RENT ($\dot{R}$)

| | Constant | $\dot{E}-1$ | $V$ | $V-1$ | $R^2$ | $D$-$W$[b] | $V^n$ |
|---|---|---|---|---|---|---|---|
| Baltimore | 8.93 | .014 | | −.67 | .36 | 1.22 | 13.3 |
| | (4.27) | (.28) | | (1.99) | | | |
| Boston | 8.75 | −.20[a] | | .95 | .66 | 2.25 | 9.2 |
| | (9.67) | (2.74) | | (2.81) | | | |
| Chicago | 5.96 | .31 | −.61 | | .63 | 1.58 | 9.8 |
| | (3.78) | (2.44) | (2.05) | | | | |
| Cincinnati | 11.76 | −.08[a] | −1.20 | | .40 | 1.30 | 9.8 |
| | (3.56) | (.62) | (2.16) | | | | |
| Cleveland | 10.04 | .18 | | −1.78 | .74 | 1.46 | 5.6 |
| | (6.20) | (2.41) | | (4.05) | | | |
| Dallas | 11.37 | .16 | | −.68 | .34 | .64 | 16.7 |
| | (3.27) | (1.68) | | (1.81) | | | |
| Denver | 8.92 | .08 | | −.61 | .17 | .32 | 14.6 |
| | (2.33) | (.45) | | (1.10) | | | |
| Houston | 16.81 | .15 | −1.18 | | .46 | 1.57 | 14.3 |
| | (2.95) | (.73) | (2.28) | | | | |
| Kansas | 10.06 | −.004[a] | −1.14 | | .51 | 1.16 | 8.8 |
| | (4.49) | (.12) | (2.70) | | | | |
| Los Angeles | 19.98 | .04 | −1.41 | −1.23 | .93 | 1.92 | 7.6 |
| | (12.37) | (.85) | (4.47) | (3.62) | | | |
| Milwaukee | 5.79 | −.03[a] | | −.25 | .17 | 1.32 | 23.2 |
| | (8.21) | (.69) | | (1.13) | | | |
| Minneapolis | 9.53 | .08 | −.87 | | .65 | 1.09 | 10.9 |
| | (5.33) | (.84) | (2.95) | | | | |
| New York | 7.57 | .002 | −.62 | −.64 | .62 | 1.75 | 6.0 |
| | (14.99) | (.08) | (2.12) | (1.93) | | | |
| Pittsburgh | 8.57 | .094 | −1.00 | | .52 | .96 | 8.6 |
| | (4.29) | (.92) | (2.13) | | | | |
| St. Louis | 6.44 | .20 | −.48 | | .64 | 2.14 | 13.4 |
| | (3.58) | (2.43) | (2.18) | | | | |
| 17-City Pool[c] | 9.08 | .05 | −.21 | −.28 | .24 | — | — |
| | (15.9) | (2.58) | (2.26) | (3.16) | | | |

*Notes:* Values shown in parentheses are the absolute value of the *t*-statistic; *D-W* is the Durbin-Watson statistic; and Percent of Change in Rent is the dependent variable.

[a] = wrong sign

[b] A Cochrane-Orcutt transformation was used for those equations exhibiting serial correlation. Since these transformed equations made the calculation of a natural vacancy rate difficult, the untransformed equations are reported here.

[c] The coefficients on the city dummy variables in alphabetical order, excluding Washington, were −1.54, −3.05, −2.60, −1.72, −3.09, .56, −1.20, 1.07, −2.82, −.68, −3.43, −1.77, −2.85, −2.59, −2.15, −2.00.

dictability" of demand fluctuations lead suppliers of rental housing to desire to maintain an inventory of unrented units. The optimal or natural vacancy rate in any market can be defined as that rate at which there is no excess demand nor excess supply and hence the rent is at its long-run equilibrium.[13] This rate is determined by the interaction of optimal search procedures on the part of households searching for dwelling space and landlords searching for tenants,[14] and by tenant turnover and institutional and market characteristics specific to each city.

Landlords face an optimizing problem in which they seek to maximize net rents through setting the gross rent and accepting the level of vacancies that rent implies. The landlord, by setting gross rent, is sampling a probability distribution of potential demanders. Since demand is unpredictable, it pays for the landlord to hold some units vacant to satisfy this fluctuating demand. Since the price structure of competing landlords and set of reservation prices for tenants are not known with precision, landlords often adjust rents on units as they become vacant, and only raise the rents for existing tenants once the new rent levels have become pretested.[15] Since rents for existing tenants can often be raised only on the expiration of their tenancy agreements, a very complex lag structure is built into the pricing behavior of landlords and hence into the relationship between optimal rents and vacancies. Optimal vacancies clearly serve as a buffer stock which prevents the need for continuous rent adjustment in a market with demand fluctuations.

---

[13] See Smith, p. 481.

[14] For a discussion of this process in the resale housing market, see Peter Chinloy (1980).

Households face a comparable optimizing problem in which they seek to minimize their shelter costs per unit quality of housing services, where shelter costs consist not only of the rental costs but also the transactions and search costs incurred in obtaining the dwellings. Since housing units are quite heterogeneous in their characteristics and are location specific, potential tenants are forced to devote considerable time to the search process in order to acquire information as to the desired bundle of housing attributes and their appropriate value. This search time increases with the size and complexity of the market, and hence search time and the natural vacancy rate are likely to increase with the size and heterogeneity of the market. The degree of market heterogeneity may be reflected in a variety of indicators including the degree of racial segmentation and the dispersion in rents.

In addition to the effect of landlord and household search behavior, the natural vacancy rate in any market will increase with the mobility or market turnover of tenants, and will be higher in areas of rapid growth where there is relatively more building ahead of demand than in areas of slow growth.[16]

An estimate of the natural vacancy rate for each city can be determined from the empirical results in Table 1. If we assume the appropriate estimating specification of equation (4) would exclude an intercept, and that $V^n$ is constant over the estimating period for each city, then the intercept in estimating equation (5) for each city can be interpreted as[17] $b_0 = b_2 V^n$, and the natural vacancy rate for each city is $b_0/b_2$. Assuming this specification, the last column in Table 1 sets out the estimated natural vacancy rate for each city.[18]

This column shows a large variation in the natural vacancy rate between cities, with Cleveland and New York having the lowest natural vacancy rates at 5.5 and 6.0 percent, respectively, and Dallas, Denver, and Houston having the highest natural vacancy rates at 16.7, 14.6, and 14.3 percent, respectively.[19] The median vacancy rate was 9.8 percent. Although these rates appear high, they are reasonable since our measure of vacancies, the difference between the gross potential total income and the actual income received, generates higher vacancy measures than the more common measure of the number of units vacant at any point in time. This happens because our measure includes bad debts, uncollectable rents, rental concessions and frictional rental losses associated with short-period rental losses or rental commissions granted on turnover, while these are rarely reflected in the common measure of vacancies.

The estimated natural vacancy rates in Table 1 can be used to examine the determinants of the natural vacancy rate on a cross-section basis. Since we are focusing on the $V^n$ between cities, we are examining the determinants of the long-run natural vacancy rate. This follows since, regardless of the adjustment lag, the fundamental differences between cities (such as zoning constraints, mobility patterns, racial composition, etc.) may be assumed to have persisted for years, and hence to have had sufficient time to have affected the market.

---

[15] This implies that tenancy discounts will exist according to length of tenure. For evidence of this, see W. A. V. Clark and Allan Heskin (1982).

[16] This point was suggested by de Leeuw and Ekanem (1971, p. 812).

[17] Alternatively, if we assume that appropriate specification has an intercept, but that the intercept is proportional to $b_2$ across cities, then $b_0/b_2$ measures the natural vacancy rate plus a given constant for each city. In the empirical work that follows, this constant would be absorbed in the intercept and the independent variables would explain the variance in $V^n$ between cities.

[18] The $V^n$ could not be determined for Philadelphia and Washington, since the model did not hold for these cities.

[19] Milwaukee is excluded in this summary because the calculated natural vacancy rate seemed unreasonably high.

The basic estimating model is set out in functional form in equation (6), with the anticipated sign indicated above each variable.[20] The expected sign of the $R$ variable is ambiguous because of its opposite influence on landlord and tenant search behavior.

$$\hat{V}^n = v\left(\overset{?}{R}, \overset{+}{RDISP}, \overset{+}{S}, \overset{+}{RS}, \overset{+}{M}, \overset{+}{\Delta SH}, \overset{-}{\Delta POP}\right), \tag{6}$$

where $\hat{V}^n$ is the estimated natural vacancy rate calculated from Table 1.[21] The variables in our estimation are as follows: $R$ is the mean level of rents, $RDISP$ is the dispersion in rents defined as the percentage standard deviation of the mean rent, $S$ is the city size in terms of population, $RS$ is the racial segmentation defined as the proportion of black and Spanish-speaking populations, $M$ is the renter mobility rate for the period, 1975–80, $\Delta SH$ is the average annual change in the total housing stock, and $\Delta POP$ is the average annual growth in population. All of these data series were derived from the 1980 Census. The model was empirically tested using $OLS$ estimation and the statistically best results are presented in Table 2.

TABLE 2—DETERMINANTS OF THE NATURAL VACANCY RATE ($V^n$)

| | Regression | |
|---|---|---|
| | (1) | (2) |
| Mobility Rate | .214 | .306 |
| (1975–80) | (1.81) | (1.95) |
| Rent Dispersion | .241 | .265 |
| | (1.97) | (2.10) |
| Average Rent | $-.139$ | $-.150$ |
| (1980) | (2.23) | (2.33) |
| Change in Housing Stock | $9.38E-06$ | $1.56E-05$ |
| (1970–80) | (2.52) | (1.99) |
| Change in Population | — | $-4.8E-06$ |
| (1970–80) | | $(-.91)$ |
| Constant | 15.04 | 13.41 |
| | (2.17) | (1.86) |
| $\bar{R}^2$ | .547 | .55 |
| $n$ | 14 | 14 |

*Note:* The absolute values of the $t$-statistic are shown in parentheses. The Natural Vacancy Rate is the dependent variable.

The empirical results support the basic model used to explain the natural vacancy rate. The results indicate that the natural vacancy rate is higher in areas that experience a higher degree of turnover. This is as anticipated since the higher the turnover rate, the higher the proportion of units being vacated and households searching for new shelter space. Consequently, the higher is the vacancy rate required and desired to allow the searching procedures to proceed efficiently.

The greater the dispersion in rents, the longer the anticipated search, and hence the higher the natural vacancy rate. The negative coefficient on the rent level indicates that the effect of higher rents on landlord opportunity costs exceeds the effect of higher rents increasing the expected gain for tenants from continuing search.

The significant positive coefficient on the growth of the housing stock variable supports the de Leeuw and Ekanem hypothesis that natural vacancy rates will be higher in areas of rapid construction. The negative but insignificant sign on the population growth variable suggests that more rapid population growth reduces the natural vacancy rate once the higher rate of construction is taken into account.

Various variable specifications were attempted to capture a racial segmentation effect, including the proportion of black to white households, and black and hispanic to white households, but the variable was al-

---

[20] Since this is a cross-sectional model, the rate of interest was not included as an explanatory variable. However, the interest rate could have a significant influence on the cost of vacancies, and hence the natural vacancy rate, over time.

[21] Philadelphia, Milwaukee, and Washington were excluded since a reasonable natural vacancy rate could not be calculated.

ways insignificant. This suggests that segmenting the market may not affect search time, possibly because the time required to search a given portion of the available market is reduced as a result of segmentation restricting the available universe, which offsets the effect of segmentation on restricting choice. Similarly, various measures of city size (population, total housing stock, rental housing stock) were also insignificant in determining the natural vacancy rate, suggesting that city size exerted neither economies or diseconomies in the search procedure.

## III. Summary

The preceeding analysis confirms the view that within some critical zone of occupancy, variations in the vacancy rate around some natural rate of vacancy exert a significant influence on the rate of change of the price of rental housing services. This result contradicts some recent studies and demonstrates that the price adjustment in rental housing is in fact sensitive to excess demand and supply conditions in the housing market. Variations in the actual vacancy rate were shown to be significant in determining the percentage change in rents at the 95 percent level in thirteen of the seventeen cities examined (and at the 90 percent level in fifteen of the seventeen cities), and in the pooled cross-section, time-series regression over all seventeen cities. The percentage change in operating expenses was also shown to be significant in determining the percentage increase in rent in most cities and in the pooled regression.

The empirical results were also used to calculate the natural vacancy rate for those cities for which the model applies. These percentage rates showed a large variation ranging from lows of 5.5 and 6.0 for Cleveland and New York, respectively, to highs of 16.7, 14.6, and 14.3 for Dallas, Denver, and Houston, respectively. The variation in the natural vacancy rate across cities was explained by a search model, interacting the search behavior of landlords for tenants and tenants for dwelling space, and by the turnover and growth rates specific to each city. The wide but predictable variation in the natural vacancy rate confirms the hypothesis that much of the variation in vacancy rates between cities reflects differences in the natural vacancy rates rather than in the degrees of market tightness, and that variations in the actual vacancy rate from its long-run normal is the appropriate variable for explaining the price-adjustment mechanism for rental housing markets.

## REFERENCES

**Alchian, Armen, A.,** "Information Costs, Pricing and Resource Unemployment," in E. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory,* New York: W. W. Norton & Co., 1970, 27–52.

**Blank, David M. and Winnick, Louis,** "The Structure of the Housing Market," *Quarterly Journal of Economics,* May 1953, *67,* 181–203.

**Chinloy, Peter T.,** "An Empirical Model of the Market for Resale Homes," *Journal of Urban Economics,* May 1980, *7,* 279–92.

**Clark, W. A. V. and Heskin, Allan D.,** "The Impact of Rent Controls on Tenure Discounts and Residential Mobility," *Land Economics,* February 1982, *58,* 109–17.

**de Leeuw, Frank and Ekanem, Nkanta F.,** "The Supply of Rental Housing," *American Economic Review,* December 1971, *61,* 806–17.

_____ **and** _____, "Time Lags in the Rental Housing Market," *Urban Studies,* February 1973, *10,* 39–68.

**Diamond, Douglas B., Jr.,** "A Note on Inflation and Relative Tenure Prices," *Journal of the American Real Estate and Urban Economics Association,* Winter 1978, *6,* 438–50.

**Eubank, Arthur A., Jr. and Sirmans, C. F.,** "The Price Adjustment Mechanism for Rental Housing in the United States," *Quarterly Journal of Economics,* February 1979, *93,* 163–83.

**Fair, Ray C.,** "Disequilibrium in Housing Models," *Journal of Finance,* May 1972, *27,* 207–21.

**Gordon, Robert A.,** "Population Growth, Housing and the Capital Coefficient," *American Economic Review,* June 1956, *46,* 307–22.

**Hendershott, Patric H. and Schilling, James D.,** "The Economics of Tenure Choice," in C. F. Sirmans, ed., *Research in Real Estate*, Greenwich: JAI Press, 1981.

**Jaffee, Dwight and Rosen, Kenneth T.,** "Mortgage Credit Availability and Residential Construction," *Brookings Papers on Economic Activity*, *2*: 1979, 333–76.

**Lowry, Ira S.,** "Rental Housing in the 1970s: Searching for the Crisis" in J. C. Weicher et al., eds., *Rental Housing: Is There a Crisis*, Washington: The Urban Institute, 1981, 23–38.

**Maisel, Sherman T.,** "A Theory of Fluctuations in Residential Construction Starts," *American Economic Review*, June 1963, *53*, 359–83.

**Muth, Richard F.,** "The Demand for Non-Family Housing," in A. C. Harberger, ed., *The Demand for Durable Goods*, Chicago: University of Chicago Press, 1960, 29–96.

**Olsen, Edgar O.,** "A Competitive Theory of the Housing Market," *American Economic Review*, September 1969, *59*, 612–21.

**Rosen, Harvey S. and Rosen, Kenneth T.,** "Federal Taxes and Homeownership: Evidence from Time Series," *Journal of Political Economy*, February 1980, *88*, 59–75.

**Smith, Lawrence B.,** "A Note on the Price Adjustment Mechanism for Rental Housing," *American Economic Review*, June 1974, *64*, 478–81.

**Institute of Real Estate Management,** *Annual Income/Expense Analysis for Apartments*, Chicago: National Association of Realtors, various years.

# On Commons and Tragedies

*By* RICHARD CORNES AND TODD SANDLER*

A wide range of market failures can be classified as "problems of the commons," in which a scarce fixed factor is not imputed a rent. Common property examples include, among others, fisheries, oil pools, hunting grounds, and the resources of outer space. Traditionally, common property analyses demonstrate the overexploitation of the scarce fixed resource; the *average product* of the variable input, not its *marginal product*, is equated to the input's rental rate when access is free and the number of exploiters is large.[1] At such a Nash equilibrium, rents of the variable factor drop to zero which *limits* the commons' overuse (Robert Haveman, 1973).

The purpose here is to extend the study of the commons to allow for a nonzero conjectural variation regarding what one exploiter thinks will be the effect of his exploitation on the exploitation efforts of others. Thus, non-Nash behavior is applied. The resulting analysis indicates that the degree of over- or underexploitation of the commons depends on the particular conjecture and the underlying technology. For a representative commons model, we show that *consistent conjectures*, those that conform with reality, cannot characterize standard Nash equilibria. Additionally, non-Nash behavior is shown to be inconsistent in all but one case, a case corresponding to zero profits even in the instance of a finite number of firms.

In Section I, Nash and Pareto equilibria for the commons are contrasted. Non-Nash equilibria are examined in Section II, and the influence of group size is investigated in Section III. The consistency of expectations and conclusions are contained in Sections IV and V.

## I. Isoprofit Curves: Nash and Pareto Equilibria

The industry consists of profit-maximizing firms, each of which has access to an exogenously fixed, common property resource. Each firm combines a single private input to the services of the commons so as to produce a single output. Henceforth, we shall refer to common property as a fishing ground, the inputs as fishing vessels, and the output as fish. With the size of the fishing ground given, the total catch, $C$, depends solely on the size of the total fleet, $R$. The implied production function exhibits diminishing returns to vessels:

$$C = F(R), F'(R) > 0, F''(R) < 0. \tag{1}$$

Both fish and vessels command an exogenously fixed price: the price of fish is unity, while that of a vessel is $p$.

Within this setting, we now model the environment of an individual firm. Its output and input levels are denoted by $c$ and $r$, respectively, and we assume that the boundary of its production possibility set is defined by a twice-differentiable function:

$$c = f(r, \tilde{R}), \tag{2}$$

where $\tilde{R} = R - r$ is the size of the rest of the fleet which plies the fishing ground. Equation (2) is appropriate for the commons, since it highlights the dependence of an individual firm's production on the actions of the rest of the industry.[2]

Most discussions of the commons assume that each firm's share of total output equals

---

*Department of Economics, The Faculties, Australian National University, Box 4, Canberra, ACT 2600, and Department of Economics, University of Wyoming, Box 3985, Laramie, WY 82071. An earlier version of this article was written while Sandler was a Visiting Fellow at the Australian National University. While assuming sole responsibility for any remaining shortcomings, we acknowledge the helpful comments of a referee.

[1] See, for example, Partha Dasgupta and Geoffrey Heal (1979, p. 61), Martin Weitzman (1974, p. 225), and Robert Dorfman (1974).

[2] On the production side, the model here is similar to that used in Dasgupta and Heal (ch. 3) and in Weitzman.

its share of the total employment of the private input. This imposes additional structure on $f(r,\tilde{R})$, which becomes

$$(3)\quad f(r,\tilde{R}) = [r/(r+\tilde{R})]F(r+\tilde{R}) = (r/R)F(R).$$

This formulation is consistent with the fish being evenly distributed throughout the fishing grounds, so that each vessel makes the same catch. One may think of this as the case of "pure" common property.

We will continue to refer to the firm's technology in terms of the general function $f(r,\tilde{R})$; however, at two points, we appeal explicitly to the structure embodied in (3). For some common property problems, alternative production structures may be more appropriate.

The Pareto optimal solution for the commons is found by choosing the industry fleet size that maximizes total industry profit, $\Pi$. The maximand is

$$(4)\qquad \Pi(R) = F(R) - pR.$$

The optimal value of $R$, $R^*$, is uniquely determined by the first-order condition, $F'(R) = p$, being *independent* of the distribution of catch between firms.

This solution is illustrated in Figure 1 where isoprofit contours, $\pi_1\pi_1$ and $\pi_2\pi_2$, show the combinations of $r$ and $\tilde{R}$ corresponding to a constant profit level for the firm.[3] Lower isoprofit contours represent greater profit levels since for a given $r$ value, $\tilde{R}$ is less; thus, profits of $\pi_1\pi_1 >$ profits of $\pi_2\pi_2$. The 45° locus, $PP$, depicts the Pareto optimal points, showing all combinations $r$ and $\tilde{R}$ whose sum is $R^*$.

The Nash equilibria correspond to those $r^*$ where the individual firm maximizes its profits while treating $\tilde{R}$ as given. The firm's problem, then, is to

$$(5)\qquad \underset{(r)}{\operatorname{Max}}\{f(r,\tilde{R}) - pr\}.$$

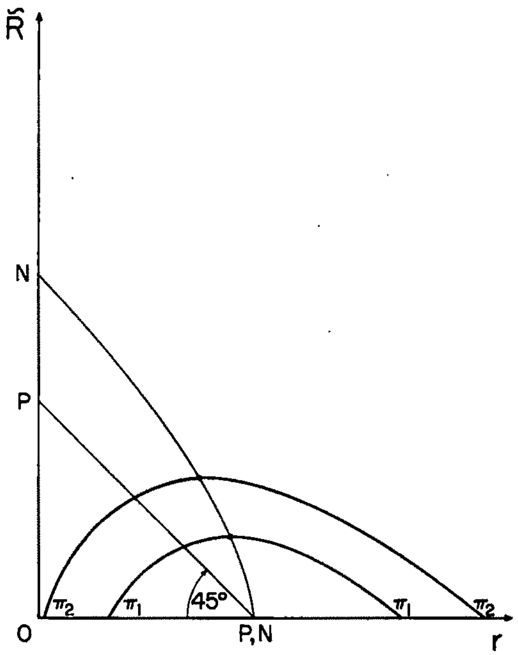


FIGURE 1. NASH AND PARETO EQUILIBRIA

This is accomplished by hiring vessels up to the point at which $p = f_r(r,\tilde{R})$. In Figure 1, the Nash solutions, generated by varying $\tilde{R}$ parametrically, lie along $NN$ which connects the maxima of the isoprofit contours. This follows because the isoprofit contour's slope is $(p - f_r)/f_{\tilde{R}}$. The slope of the Nash reaction path is, however,

$$(6)\qquad d\tilde{R}/dr = -f_{rr}/f_{r\tilde{R}}.$$

Furthermore, when $\tilde{R} = 0$, Nash behavior implies the same level of $r^*$ as does industry profit maximization; that is, under monopoly, the firm is the industry. Thus, points $P$ and $N$ correspond to one another on the horizontal axis. If the firm's production function is that in (3), then the slope of the Nash reaction curve is greater in absolute terms than that of the industry profit-maximizing locus $PP$. To demonstrate this, differentiate (3) to give

$$f_r = F'(R) + (\tilde{R}/R)[F(R)/R - F'(R)].$$

Recalling that $R = \tilde{R} + r$, it immediately follows that

---

[3] We should note at the outset that the profit function $\pi(r,\tilde{R})$ is not necessarily quasi concave. However, further restrictions may be imposed which are sufficient for quasi concavity, and these we impose.

$$f_{r\tilde{R}} = f_{rr} + \frac{[F(R)/R - F'(R)]}{R} > f_{rr}.$$

The derivative $f_{r\tilde{R}}$, then, is a numerically smaller negative number than $f_{rr}$, so that at any point on the Nash reaction curve $|d\tilde{R}/dr| > 1$.

## II. Hybrid or Non-Nash Behavior

### A. *A General Formulation*

By hybrid behavior we refer to maximizing behavior predicated on conjectures that one exploiter holds with respect to the way in which the other exploiters will respond to his own fishing efforts. In the standard common property model, these conjectures are absent; the other exploiters' level of fishing activity is taken as given (or exogenous) and independent of one's own maximizing activity. The existence of a positively (negatively) sloped Nash reaction path shows, however, that such a zero conjectural variation assumption is not consistent with reality.

A general formulation for hybrid behavior is

$$(7) \qquad d\tilde{R}^e/dr = g(\theta, r),$$

where $\theta$ is a parameter (or vector of parameters), representing the other exploiters' (anticipated) responsiveness to the exploiter's own fishing activity. In (7), the conjectured response depends on the exploiter's fleet size. Generally, (7) can be integrated to give an endogenous relationship for $\tilde{R}^e$:

$$(8) \qquad \tilde{R}^e = \tilde{R}^e(\theta, r, K),$$

where $K$ is a constant of integration, dependent on the rest of the fishing firms' initial fleet size. Other representations for the conjectural variations are possible, provided that the implied differential equation is integrable; thus, $\tilde{R}$ itself may be an argument.

The introduction of anticipations does not alter the Pareto optimality conditions, since presumably the joint profit-maximizing assumption already internalizes all expectations. In contrast, the introduction of nonzero conjectural variations affects the first- and second-order conditions for the firm's profit-maximizing behavior. The firm now chooses its fleet size to

$$\underset{\{r\}}{\text{Max}}\{f[r, \tilde{R}^e(\theta, r, K)] - pr\},$$

which depends on its own fleet size and the conjectured fleet size for the rest of the firms. The first-order condition[4] requires

$$(9) \qquad f_r + f_{\tilde{R}^e}\, d\tilde{R}^e/dr = p,$$

which degenerates to the Nash solution only when conjectural variations are zero. When, however, $d\tilde{R}^e/dr$ is positive (implying that one's own increased fishing effort is expected to induce others to follow suit), the firm's optimal fleet and the commons' tragedy will be less than that implied by the standard solution since $p - f_r > 0$ in (9). For negative conjectures, the firm's optimal fleet exceeds those of the standard solution; the tragedy of the commons is intensified.

The above analysis and solution are displayed for the representative firm in Figure 2, where three isoprofit curves are shown. Curves $K_1\tilde{R}^e$, $K_2\tilde{R}^e$, and $K_3\tilde{R}^e$ represent three endogenous paths for $\tilde{R}^e$, each of which differs by the initial fleet size for the rest of the fishing firms. The slopes of these contours, $d\tilde{R}^e/dr$, are the conjectural variations given in (7). The locus of tangencies between the isoprofit curves and the expectations contours trace out a hybrid reaction path, *HH*, for the representative firm.[5]

The dashed path, *H'H'*, is the hybrid path when conjectures are negative; the repre-

---

[4] If the first-order condition is to represent optimal behavior, a second-order condition must also hold. This latter condition requires that, in the neighborhood of the point characterized by (9), that the expectations contour described in (8) and shown in Figure 2 have an opposite convexity to that of the isoprofit curves *or*, in the case of identical kinds of convexity, that the expectations contours have less curvature than the isoprofit curves.

[5] This follows from a rearrangement of (9) to give

$$(p - f_r)/f_{\tilde{R}^e} = d\tilde{R}^e/dr,$$

where the two sides correspond to the slope of the isoprofit and expectations contours, respectively.

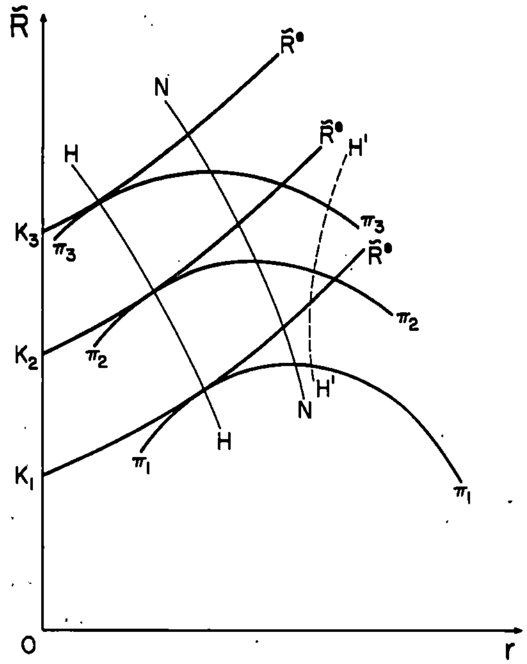


FIGURE 2. HYBRID REACTION PATHS

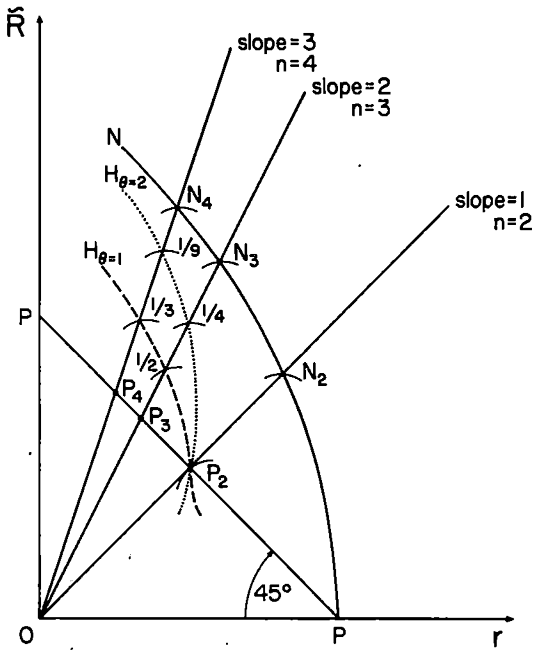


FIGURE 3. IDENTICAL FIRMS $d\tilde{R}^e/dr = (n-1)^{-\theta}$

sentative firm expects others to decrease their activity in response to its own activity. As shown in Figure 2, the hybrid reaction path for positive conjectures lies to left of the Nash path, while that for negative conjectures lies to the right. The hybrid path's relative position vis-à-vis the Pareto path depends on both the curvature (hence third-order partials) of the profit surface and the shape of the expectations contours; no general statement is possible at this juncture.

### B. *A Specific Illustration*

An interesting variety of behaviors can be displayed with the following specific formulation of the conjectural variation:

$$d\tilde{R}^e/dr = (r/\tilde{R})^{\theta}. \tag{10}$$

In (10), the firm anticipates that its increased fishing efforts will induce others to follow; in particular, the larger is the industry's size relative to that of the firm, the smaller is the anticipated response. When there are $n$ identical firms, the formulation in (10) becomes

$$d\tilde{R}^e/dr = (n-1)^{-\theta}, \tag{11}$$

thus permitting industry size, $n$, to be related to suboptimality. The first-order condition implied by (11) is

$$(p-f_r)/f_{\tilde{R}^e} = (n-1)^{-\theta}. \tag{12}$$

Figure 3 illustrates the specific formulation for three different values of $\theta$. Equilibria can be found as the number of firms varies by focusing on tangencies (for Pareto points) and maxima (for Nash points) along different rays from the origin. For example, a ray with a slope of 1 corresponds to $n = 2$, since $\tilde{R}/r = 1$; a ray with a slope of 2 relates to $n = 3$, since $\tilde{R}/r = 2$; and so on. Pareto optimal points for different industry sizes lie along $PP$, whereas Nash equilibria are on $NP$. Subscripts indicate group size. The ratio $ON_i/OP_i$ expresses an "index of tragedy": ratios greater than 1 correspond to a misallocation, where large ratios imply greater inefficiency.

Hybrid equilibria can also be found along these rays. When $\theta = 1$, the dashed path in Figure 3 is the hybrid path. Points along it satisfy (12), and hence refer to tangencies between the expectation contour and the isoprofit curve. The hybrid equilibrium for $n = 3$, for example, corresponds to the lowest isoprofit curve with slope 1/2 on the ray with slope 2. The dotted path is the hybrid path for $\theta = 2$ and displays a faster convergence to Nash behavior as industry size grows than that associated with the dashed path. This formulation contains the following implications: 1) values of $\theta \geqslant 1$ imply a convergence to Nash behavior as industry size grows; 2) the larger is $\theta$ (provided that $\theta > 1$) the faster is this convergence; and 3) $\theta = -1$ indicates Pareto optimal behavior, regardless of industry size.

## III. On the Influence of Group Size

Using the production structure of (3) we can analytically display the influence of group size for the symmetric case. Each firm now chooses its fleet size to

$$\underset{\{r\}}{\mathrm{Max}}\{r/[r+\tilde{R}^e(\theta,r,K)]\}\cdot\{F[r+\tilde{R}^e(\cdot)]-pr\}.$$

The first-order condition requires

$$(13)\quad F(R)/R-(1+d\tilde{R}^e/dr)/n \times[F(R)/R-F'(R)]=p,$$

where $R = r + \tilde{R}^e$ represents the anticipated catch of the commons. If, for example, $d\tilde{R}^e/dr = 0$, then as $n$ goes to infinity, the average product of vessels is equated to vessel rental rates; the model then generates the traditional Nash solution. This conventional result is clearly seen to depend on atomistic competition and would not necessarily characterize commons shared by a finite number of firms.

When $d\tilde{R}^e/dr = n-1$, so that a firm fully anticipates that others will act as they do, (13) indicates that the (industry's) vessels' marginal product will then be equated to the rental rate (i.e., $F'(R) = p$). Pareto optimality results since the effects on the unpaid input (the commons) is fully internalized through anticipations.

For hybrid behavior, (13) shows that the influence of anticipations may dissipate rapidly when industry size increases. Such is the case whenever $(1+d\tilde{R}^e/dr)$ increases at a slower rate than $n$ as $n$ gets large (for example, $d\tilde{R}^e/dr = n^\alpha, \alpha < 1$). When, however, expectations are a linear function of $n$, as they are when $d\tilde{R}^e/dr = n-1$, the influence of anticipations do not diminish with group size. In the specific case presented in Section II, increases in group size reinforces the reduced effects of anticipations as $n$ increases whenever $\theta > 1$. If $n$ is small, anticipations can be quite crucial in determining equilibrium.

## IV. Equilibrium and Consistent Conjectures

For Nash and non-Nash behavior, conjectures are consistent when they conform to reality. We follow the oligopoly literature (for example, Timothy Bresnahan, 1981; and Martin Perry, 1982) by defining a conjectural variation as consistent if it is identical to the optimal (actual) response of the other agent(s) in a neighborhood of the equilibrium based upon the conjecture. Thus, for two firms, a conjecture is consistent when, at the equilibrium, the slope of each firm's reaction path equals the corresponding conjectural variation held by the other firms.

In the model of Section I, it turns out that a Nash equilibrium *cannot be consistent.* Consistency requires that the conjectured response of the representative agent—which is zero in the Nash model—should equal his actual response. In Figure 1, this means that Nash reaction curve *NN* should be vertical in the neighborhood of equilibrium. But we know from (6) that this can only happen if $f_{r\tilde{R}} = 0$. So long as there is interdependence, which typifies the commons, the Nash reaction must have a finite slope.

One might hope that when the class of models is enlarged by admitting nonzero conjectural variations, a consistent conjectures equilibrium becomes a possibility. Interestingly enough, one can show that such an equilibrium fails to exist even within this broader class of models of common property for any value of $d\tilde{R}^e/dr$, except for minus one. This is demonstrated elsewhere[6] for symmetric equilibrium by setting the conjectural variation, $\sigma = dR^e/dr$, equal to the slope of the hybrid reaction path, and then showing that $\sigma$ must equal $-1$.

What other features does such a consistent equilibrium possess? To begin with, we note that so long as the average product exceeds price, so that positive profits are made by existing firms, such an equilibrium cannot be sustained. If an individual firm expects the rest of the industry to reduce its input by one unit in response to a one unit increase in the firm's input, so that industry output will remain unchanged, then it will pay the firm to increase its fleet size, since its profit is

$$\pi(r,R)=(r/R)F(R)-pr.$$

Therefore, the effect on its profit in response to its own fleet size is

$$\partial\pi/\partial r = F(R)/R-p.$$

Thus, the consistent conjectural equilibrium requires zero profits. This is particularly interesting since it is relevant in the competitive case in which, in the absence of fixed costs and barriers to entry, additional firms will be persuaded to exploit the common property up to that point where average product equals price and profits are eliminated.

## V. Conclusions

This article has introduced oligopoly-like reactions into the analysis of the commons by allowing agents to perceive an interdependence between their actions and those of others. This nonzero conjectural variation is crucial in determining the degree of misallocation associated with the commons. Our analysis of conjectural consistency has led to an interesting conclusion: if there is more than one firm exploiting the commons, consistency involves exploitation up to the point where profits are zero. This is so even if entry is restricted to a finite number of firms. The Nash equilibrium, while it suggests overexploitation of the resource, generates a less pessimistic prediction than does the consistent conjectures equilibrium. In this latter context, two is a large number.

---

[6] The proof is in a longer version of this article, available upon request from either author.

## REFERENCES

**Bresnahan, Timothy,** "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, *71*, 934–45.

**Dasgupta, Partha and Heal, Geoffrey,** *Economic Theory and Exhaustible Resources*, Cambridge: Cambridge University Press, 1979.

**Dorfman, Robert,** "The Technical Basis for Decision Making," in E. T. Haefele, ed., *The Governance of Common Property Resources*, Baltimore: Johns Hopkins University Press, 1974, 5–25.

**Haveman, Robert,** "Common Property, Congestion, and Environmental Pollution," *Quarterly Journal of Economics*, May 1973, *7*, 278–87.

**Perry, Martin,** "Oligopoly and Consistent Conjectural Variations," *Bell Journal of Economics*, Spring 1982, *13*, 197–205.

**Weitzman, Martin,** "Free Access vs. Private Ownership as Alternative Systems for Managing Common Property," *Journal of Economic Theory*, June 1974, *8*, 225–34.

# Frameworks for Analyzing the Effects of Risk and Environmental Regulations on Productivity

*By* W. KIP VISCUSI*

Although regulatory agencies have promulgated health, safety, and environmental regulations largely on the basis of their presumed benefits, the adverse economic effects of these policies are becoming an increasingly prominent concern. A series of studies has linked these regulations to the productivity slowdown, inflation, and unemployment.[1] Significant economic effects of this type should be expected since the present value of the costs associated with major risk and environmental regulations proposed between 1975 and 1980 was \$332.2–\$846.5 billion.[2] These cost impacts are both quite large and highly variable. My concern in this paper is with how both the level of these costs and their uncertainty affect productivity.[3]

The existence of a negative relationship between the regulatory burden and capital investments, and consequently productivity, is not controversial. A conventional model of this type is developed in Section I. If, however, these regulations change over time and firms' investment decisions are irreversible, there will be additional distortions, as shown in Section II. In Section III, I show that uncertainty regarding these regulatory changes exacerbates the adverse productivity effects even for risk-neutral firms.

## I. Regulatory Impacts with Reversible Investments

If firms' investments are completely reversible, we can treat the firm as if it were renting capital on a period-by-period basis. This can be done using either a single period model, or, in a multiperiod context, by assuming that the firm is myopic. In this section, I develop a fairly conventional model of this type to establish a point of reference for the subsequent results. Consider firms with two choice variables, the output level $q$ and the quality level $z$, which will be scaled without loss of generality so that $z$ is in the interval $[0,1]$, where higher values of $z$ represent higher quality levels. For risk regulations, the value of $1-z$ represents the risk level per unit of output, such as the product failure probability or the injury rate per worker.[4] For environmental situations, $1-z$ can be viewed as the level of pollution per unit of output.

The firm sells its output at a price $v$, and incurs production costs $C(q)$, where $C(q) \geqslant 0$, $C' > 0$, and $C'' > 0$. Raising the quality level $z$ is also costly, with the unit cost of producing quality $z$ being given by $G(z)$, where $G(z) \geqslant 0$, $G' > 0$, and $G'' > 0$. Finally, firms are penalized $x$ per unit of risk or pollution, so that the overall regulatory penalty is $xq(1-z)$. The value of $x$ represents the total quality-related penalty, including rewards transmitted through market forces, but for concreteness I will treat the entire value of $x$ as a policy parameter. If $x$ reflects the value society places on each unit of quality, the regulatory policy will lead to an efficient outcome. The financial incentives imposed by workers' compensation systems or by proposed pollution tax policies best fit this structure. The model is also applicable to standards policies for which the penalties are linked to the extent of noncompliance once we reinterpret $z$ as the degree of regulatory compliance.[5]

---

*Professor and Director, Center for Study of Business Regulation, Fuqua School of Business, Duke University, Durham, NC 27706. This research was supported in part by the Center for Study of Business Regulation.

[1] These studies include Edward Denison (1978), Robert Haveman and Gregory Christainsen (1980) Marvin Kosters (1980), Paul MacAvoy (1979), Peter Pashigian (1981), and Paul Portney (1980).

[2] I present this estimate in my 1983 study.

[3] The role of regulatory uncertainty is often stressed by businessmen. As George P. Shultz observed, it introduces "a real wild card in investment decisions" (1980, p. 14).

[4] In the case of job safety, one can easily modify the formulation below so that the firm picks the number of workers $L$ rather than the output level $q$.

Firms consequently select the level of $q$ and $z$ to maximize profits $\pi^0$, or they

$$\max_{q,z} \pi^0 = vq - C(q) - qG(z) - xq(1-z), \tag{1}$$

yielding first-order conditions that reduce to[6]

$$C' = v - G(z) - x(1-z); \tag{2}$$

$$G' = x. \tag{3}$$

For given quality level $z$, the regulation lowers the marginal benefits of production (or raises the marginal costs if one rearranges equation (2)), lowering the optimal output level. As equation (3) implies, the optimal quality level is set where its unit cost equals the regulatory penalty $x$. Finally, more stringent regulatory penalties will raise the overall quality level and depress output, as

$$\partial z/\partial x = 1/G'' > 0$$

and $$\partial q/\partial x = -(1-z)/C'' < 0.$$

## II. Irreversible Decisions and Changes in Regulatory Policies

If there are no impediments to changes in the firm's decisions, shifts in regulatory policies can be assessed using a one-period model as in Section I. Firms usually have much more limited discretion. Modifications in plant capacity, particularly in the downward direction, are typically quite costly due to the specialized nature of the capital stock, transactions costs involved in its resale, and the absence of a strong resale market. Similarly, once a firm has complied with risk and environmental regulations by, for example, installing a ventilation system to control hazardous fumes, the costs of compliance represent a sunk cost that cannot be readily recovered.

Although the degree of inflexibility spans a continuum of possibilities, both in the upward and downward directions, I will focus on irreversibilities, where there is complete downward inflexibility in the enterprise's decision. Under this formulation, which was introduced by Kenneth Arrow (1968), the firm cannot reduce past capital investments, but it is free to augment these investments. This asymmetric approach is intended to be an approximation to the greater difficulties associated with decreasing past investments. While these investments may not be characterized by complete downward inflexibility, as the transactions costs of disinvestment increase, the results approach those under strict irreversibility.

Consider firms making decisions in a two-period model. Since in a single period, $C(q)$ and $G(z)$ represented the capital stock rental costs for one period, I will treat the role of irreversibility as making the first-period choice of $q$ and $z$ tantamount to a rental commitment that it cannot reduce.[7] If we denote the two periods by subscripts $i$, where $i = 1,2$, the irreversibility assumption is that $q_2 \geqslant q_1$ and $z_2 \geqslant z_1$. Once an enterprise selects an output or quality level, it cannot reduce it.

### A. *Increasing Regulatory Stringency*

Suppose that firms are aware that the regulatory penalty will rise from $x$ in period 1 to $x^*$ in period 2, where $x^* > x$. Regulatory policies that are phased in over several years have this character, as policymakers attempt to ease the adjustment burdens on the affected industries. Increasing the regulatory penalty will enhance the attractiveness of quality investments and will diminish the incentive to make capacity investments by raising the regulatory burden per unit of output. Since the firm cannot reduce its output level, it will remain at whatever level is selected in period 1, or

---

[5] The only additional complication is that corner solutions play a more important role in the case of standards. For firms that choose complete compliance, the penalty term involving $z$ drops out of the analysis, but the unit cost terms remain, where an optimizing firm will set $z$ at the minimal level to ensure compliance in this instance.

[6] The second-order conditions are also satisfied since $\pi^0_{qq} < 0$, $\pi^0_{zz} < 0$, and $\pi^0_{qq}\pi^0_{zz} - (\pi^0_{qz})^2 = qC''G' > 0$.

[7] In a two-period model, this approach gives much more realistic results than examining actual capital investments since investments in the first period will have a longer useful life. In addition, as in the case of Arrow's analysis (p. 5), I will abstract from the influence of depreciation since it plays no essential role.

$$(4)\qquad q_2 = q_1.$$

An optimizing firm will, however, raise $z$ in the second period. The cost functions remain the same as before, but I will sometimes use the notation $G_i$ to denote $G(z_i)$, and similarly $C_i$ for $C(q_i)$.

I will follow the usual dynamic programming approach of considering the decisions in the second period, given the first-period choices, and then work backwards to analyze the optimal first-period choices. The value of $z_2$ is picked to maximize the profits $\pi^*$ in the second period, or

$$\underset{z_2}{\mathrm{Max}}\, \pi^* = vq_1 - C(q_1) - q_1G(z_2) - x^*q_1(1-z_2),$$

yielding

$$(5)\qquad G_2' = x^*.$$

This equation, which is the second-period analogue of equation (3), indicates the optimal $z_2$ level as a function of $x^*$. Equation (5) thus implicitly defines a function $Z(x^*)$, which gives the optimal value of $z_2$ as a function of $x^*$, where $Z' > 0$. It is noteworthy that $z_2$ depends only on $x^*$ and not on any other parameters or on any decisions in period 1. The optimal decisions in period 2 are defined by equations (4) and (5).

The task in period 1 is to maximize discounted expected profits $\hat{\pi}$, where the discount factor $\beta$ is the inverse of one plus the interest rate. The values of $q_1$ and $z_1$ are selected to

$$\underset{q_1, z_1}{\mathrm{Max}}\, \hat{\pi} = vq_1 - C(q_1) - q_1G(z_1) - xq_1(1-z_1) + \beta[vq_1 - C(q_1) - q_1G(Z) - x^*q_1(1-Z)],$$

leading to first-order conditions that can be rewritten as

$$(6)\qquad G_1' = x;$$

$$(7)\qquad C_1' = v - \frac{G_1 + x(1-z_1) + \beta G_2 + \beta x^*(1-Z)}{1+\beta}.$$

As regulations become more stringent, ideally firms would like to expand their quality investment and diminish their capacity investment. The increase in the quality investment is feasible, as firms select $z_i$ on a myopic basis in each period (equations (5) and (6)) based on the regulatory penalty and marginal investment costs. These quality investments are independent of the output level so that the first-order conditions are not mutually dependent; the $z_i$ values affect capacity, but there is no reverse linkage. Since capacity investments cannot be diminished, there is no additional investment in period 2, and the capacity level for both periods is selected in period 1 to satisfy equation (7).

The fundamental concern here is how the resulting capacity investment levels compare with those under the reversible cases. Because of the recursive nature of the results (i.e., the $z_i$ influence $q_1$ but are not affected by it), this relationship hinges only on the capacity investment first-order conditions. The condition that the terms on the right side of equation (2) exceed the right side of equation (7) reduces to

$$G(z(x)) + x(1-z(x)) < G(Z(x^*)) + x^*(1-Z(x^*)) \equiv H,$$

or showing that the quality-related costs per unit of output are greater in situations in which $x$ is increased to $x^*$ than in one-period problems. To prove that $H$ is increasing in $x$ differentiate and apply the envelope theorem, yielding

$$(8)\qquad dH/dx^* = 1 - Z > 0.$$

The prospect of tighter regulations depresses the value of $q_1$, compared with the reversible case, as the expression on the right-hand side of equation (2) always exceeds the comparable terms in equation (7).[8]

Upon total differentiation of equations (6) and (7), it can be shown that

$$\partial z_1/\partial\beta = \partial z_1/\partial x^* = 0,$$
$$\partial q_1/\partial\beta < 0, \quad \text{and} \quad \partial q_1/\partial x^* < 0.$$

The policy shift in period 2 does not alter $z_1$, but it does depress the optimal output level. An enterprise facing a schedule of increasingly stringent regulations will reduce its initial output level and be myopic in its quality investments. Increasing the discount factor places a greater weight on these second-period choices, leading to a reduction in the output level.

### B. *Decreasing Regulatory Stringency*

When regulations are loosened rather than tightened, there are many similarities in the technical aspects of the analysis, but some significant differences in the nature of the results. If the penalty falls from $x$ to $x_*$ in period 2, where $x > x_*$, the firm cannot reduce $z$, so that $z_2 = z_1$, but it can increase its output level. It will do this in period 2 to maximize second-period profits $\pi_*$, or

$$\underset{q_2}{\operatorname{Max}} \pi_* = vq_2 - C(q_2) - q_2G(z_1) - x_*q_2(1-z_1),$$

yielding

$$C_2' = v - G_1 - x_*(1-z_1). \tag{9}$$

Equation (9) implictly defines the optimal $q_2$ as a function of $x_*$ and $z_1$, where

$$\partial q_2/\partial x_* < 0;$$

and $$\partial q_2/\partial z_1 = (x_* - G_1')/C_2'' < 0.$$

This relationship will be denoted by $Q(x_*, z_1)$. The effect of $z_1$ on $q_2$ involves two competing effects. I will show below that the marginal penalty $x_*$ per unit of quality is smaller than the increased unit cost of quality $G_1'$ at the optimal $z_1$. This gap arises since the firm will overinvest in $z_1$ in period 1 relative to the quality investment it would have chosen if the penalty were always $x_*$. As a result, the high quality investment commitment from period 1 depresses output in period 2 since the marginal costs imposed by $z_1$ exceed the marginal reduction in regulatory penalties in period 2.

The decision in period 1 is to

$$\underset{q_1, z_1}{\operatorname{Max}} \pi^{**} = vq_1 - C(q_1) - q_1G(z_1) - xq_1(1-z_1) + \beta[vQ - C(Q) - QG(z_1) - x_*Q(1-z_1)],$$

leading to the first-order conditions

$$\pi_{q_1}^{**} = 0 = v - C_1' - G_1 - x(1-z_1); \tag{10}$$

$$\pi_{z_1}^{**} = 0 = -q_1G_1' + xq_1 + \beta Q(x_* - G_1'). \tag{11}$$

Equation (10) is equivalent to the results for the myopic case in equation (2), for fixed values of quality $z_1$. Comparing equations (9) and (10), it is clear that as the regulatory penalty is loosened from $x$ to $x_*$, the incentive to expand output in period 2 is increased since the only terms that change will be the choice variable $q_2$ (through $C_2'$) and the regulatory penalty per unit of output, which falls from $x(1-z_1)$ to $x_*(1-z_1)$.

Rewriting equation (11) in terms of the marginal investment cost yields

$$G_1' = (xq_1 + \beta x_*Q)/(q_1 + \beta Q) < x, \tag{12}$$

since $x_* < x$. Compared to the myopic quality choice in equation (3), the firm always underinvests in quality if the regulation will be loosened. This effect arises from a desire to prevent having overinvested in quality from the standpoint of the second period's regulatory policy in a situation in which investment decisions are irreversible.

---

[8] This result also implies that in the case of corner solutions in which the quality level is 1, there will be no effect of more stringent regulations on either the output or quality level since $1-z$ equals zero once quality hits the maximum level.

Assuming the second-order conditions are fulfilled,[9] the effect of a relaxation of the regulatory policy can be summarized as follows:

$$\frac{\partial q_1}{\partial x_*} > 0; \quad \frac{\partial q_1}{\partial \beta} < 0; \quad \frac{\partial z_1}{\partial x_*} > 0; \quad \frac{\partial z_1}{\partial \beta} < 0.$$

Higher values of $\beta$, which reflect a greater weight on second-period decisions, depress both forms of investment. The $x_*$ results differ from those for situations of increasing regulatory stringency since the second-period penalty affects $z_1$ as well as $q_1$. This difference arises because the firm no longer chooses $z_1$ myopically but must make a permanent quality investment commitment in period 1. This commitment is dependent on the penalty levels and capacity choices in both periods. Increases in the second-period $x_*$ boost the quality commitment $z_1$. Unlike the situation of increasing regulatory stringency, there is no direct effect of $x_*$ on $q_1$ since the irreversibility contraint on capacity investments is not binding, as the firm will increase its output in period 2. Since $z_1$ affects $q_1$, however, there is an indirect effect of $x_*$ on $q_1$. The higher $z_1$ values induced by increases in $x_*$ reduce the marginal regulatory penalties on output, so that increases in $x_*$ also enhance the incentives for capacity investments.

The parallels with the situation of increasing regulatory stringency are clearer if one views these policies in terms of departures from the initial policy $x$. The more $x_*$ is lowered below the initial penalty level, the more both $q_1$ and $z_1$ will be depressed. Similarly, the more $x^*$ exceeds $x$, the greater will be the reduction in $q_1$. All shifts in regulatory policy depress the initial output level. In the case of increasingly stringent regulations, it is because firms underinvest to avoid an irreversible commitment to a level of production that will subsequently be inefficiently large; in the case of regulations being relaxed it is the feedback effect of reduced quality investments that diminishes output. Increasing the weight $\beta$ on the second-period decisions always reduces both $q_1$ and $z_1$ when there is a shift in regulatory policies.

One might have expected that a known regulatory policy change would have influenced only one of the two forms of irreversible investment since a change in $x$ will only make one of the first-period irreversibility constraints binding. The actual impact is more pervasive due to the interdependence of the two types of investment when regulations are being relaxed.

## III. Uncertain Regulatory Policies

While industries occasionally face a known schedule of regulatory changes, particularly when very stringent regulations are being phased in gradually, future shifts in the regulatory policy are usually not specified in advance. This uncertainty is particularly great for regulations of hazards for which the available medical evidence is imprecise. The uncertainty regarding policy shifts, for any particular set of information, compounds these scientific uncertainties. If one assumes that firms are risk neutral, one might expect that the role of uncertainty could be treated using the results of Section II. After calculating the expected penalty $\bar{x}$ in period 2 and noting whether or not it involved a decrease or an increase in regulatory stringency, one might then apply the pertinent model for known regulatory changes. As the subsequent results will indicate, such a relationship does not hold, as regulatory uncertainties introduce new complications into the firm's decisions and the policy design process.

### A. *Implications for Firms' Decisions*

Suppose the firm is facing a binary policy lottery in period 2, where the penalty may rise to $x^*$ or fall to $x_*$, where $x^* > x > x_*$. The situation in which one of the second-period penalties equals $x$ poses no additional complications for the analysis below, but for concreteness I will assume that the strict inequalities hold. If both $x^*$ and $x_*$ equalled $x$, the analysis would be the same as in Section II. The firm assesses the probability that $x^*$ will be the penalty as $p$ and the probability that $x_*$ will be the penalty as $(1-p)$, and based on these expectations it is assumed to maximize discounted expected profits $\pi$.[10]

---

[9] This requirement is that $C_1''[G_1''q_1 - \beta Q_{z_1}(x_* - G_1') + \beta G_1''Q] - (G_1' - x)^2 > 0$, where $Q_{z_1} = (x_* - G_1')/C_2''$.

The decision in period 2 parallels the results for the situation of certain policy changes. If $x$ rises to $x^*$, there will be no change in the output level since the firm cannot reduce output below its first-period level. The firm will, however, raise its quality investment $z_2$, which is given by $Z(x^*)$ and is defined by equation (5) as before. Similarly, reductions in $x$ to $x_*$ lead the firm to make no change in its quality investment but provide an incentive to raise output to $Q(x_*, z_1)$, which is implicitly defined by equation (9). From the standpoint of the second-period decisions, the uncertainty only affects $q_2$ indirectly through its effect on $z_1$.

The first-period problem for the profit-maximizing firm is a variant on the earlier problems for the certainty case since the firm will

$$\begin{aligned}\underset{q_1, z_1}{\operatorname{Max}}\, \pi = {} & vq_1 - C(q_1) - q_1 G(z_1) - xq_1(1-z_1) \\ & + \beta p\left[vq_1 - C(q_1) - q_1 G(Z) - x^* q_1(1-Z)\right] \\ & + \beta(1-p)\left[vQ - C(Q) - QG(z_1) \right. \\ & \left. \qquad - x_* Q(1-z_1)\right],\end{aligned}$$

producing the first-order conditions

$$(13)\quad \begin{aligned}\pi_{q_1} = 0 = {} & v - C_1' - G_1 - x(1-z_1) \\ & + \beta p\left[v - C_1' - G_2 - x^*(1-Z)\right];\end{aligned}$$

$$(14)\quad \begin{aligned}\pi_{z_1} = 0 = {} & -q_1 G_1' + xq_1 \\ & + \beta(1-p)\left[Q(-G_1' + x_*)\right].\end{aligned}$$

These conditions can be rewritten as

$$(15)\quad C_1' = v - \frac{G_1 + x(1-z_1) + \beta p G_2 + \beta p x^*(1-Z)}{1+\beta p};$$

$$(16)^{11}\quad G_1' = \frac{xq_1 + \beta(1-p)x_* Q}{q_1 + \beta(1-p)Q} > x_*.$$

Unlike the results for the certainty situation, neither of the first-period decisions takes the same form as with reversible investment decisions. Uncertainty affects both first-period decisions and does not hinge on the expected penalty $\bar{x}$. Rather, the choice of output $q_1$, which will be increased if $x$ decreases and will remain the same if $x$ rises to $x^*$, is not directly affected by $x_*$, but it is dependent on $x^*$. The reason for this asymmetry is that the irreversibility constraint will only be binding if $x^*$ is the penalty so that only in this situation will $q_1$ be the output level. As indicated by equation (15), the firm equates the marginal production cost with the price less the conditional expected costs associated with low quality, where this expectation is restricted to the states in which $q_1$ is the output level. Equation (7) for the case of a certain increase in regulatory stringency is simply a special case of equation (15), where the value of $p$ equals one.

The quality choice $z_1$ also is affected by a conditional expectation, but here the firm sets the marginal cost of quality investments equal to the expected penalties weighted by the output levels, the probabilities, and the discount factor. Equation (16) represents a generalization of equation (12) for the case of decreasing regulatory stringency. It is not the expected penalty $\bar{x}$ that influences decisions since the only second-period penalty value that enters is $x_*$. The value of $x^*$ is irrelevant since a tighter penalty will lead the firm to augment its initial quality investment, so that the first-period quality commitment will no longer be a binding constraint. With irreversibilities, the focus should not be on the expected penalties, but on the level of the regulatory policy in states in which the irreversibility constraint is binding.

Equations (15) and (16) have unambiguous implications regarding the effect of future regulatory policy lotteries. Consider first the level of $z_1$ when compared to the reversible case. With myopic quality investments, the firm equates $G'$ with $x$, but in the presence of regulatory uncertainty, the marginal cost of quality investments is equated to the expression on the right side of equation (16), which is clearly less than $x$. If there is any nonzero probability that the penalty will be relaxed, firms facing regulatory uncertainty will underinvest in quality when compared with the myopic case.

---

[10] The introduction of risk aversion would reinforce the nature of the findings below by creating additional incentives for underinvestment.

[11] As in the certainty case, I assume that the second-order conditions are also fulfilled.

Uncertain regulatory policies will also lead to lower values of initial output if the right-hand side of equation (15) is below that in equation (2), a requirement that reduces to

$$(17)\quad [G_1+x(1-z_1)]+\beta p[G_2+x'(1-Z)] > [G+x(1-z)]+\beta p[G+x(1-z)].$$

In the myopic case, the optimal $z$ minimizes $G+x(1-z)$, so that any effect of uncertainty on the $z_1$ selected will raise these quality-related costs in period 1. Except when $1-p$ equals zero (see equation (16)), there is always a distortion of the choice of $z_1$. The first bracketed term on the left-hand side of equation (17) will exceed its counterpart on the right-hand side, except when $p$ equals 1, in which case the two terms are equal. From equation (8) above we know that the second bracketed term on the left-hand side of equation (17) exceeds the final term on the right-hand side except when there is no chance of a price increase (i.e., when $p=0$). Uncertain regulatory policies will depress output provided *either* there is a nonzero chance of a penalty increase or a nonzero chance of a penalty decrease. Any nontrivial regulatory policy lottery will depress output.

Greater weight on the second-period outcomes in which the policy lottery occurs also have an adverse effect. Increasing the value of $\beta$ has an unambiguous effect on $q_1$ and $z_1$. Total differentiation of equations (13) and (14) produces the result that $\partial q_1/\partial\beta<0$ and $\partial z_1/\partial\beta<0$, or greater weights on the period with uncertain regulatory policies will depress both types of investment in period 1.

Assessing the effects of changes in the lottery structure are more complicated since there are conflicting influences due to the complex interrelationships between $q_1$ and $z_1$. The direct effects of $p$ on these variables are clear cut and, for the remainder of the paper, I will assume that these direct impacts are dominant.[12] Equivalently, I will focus on the effects of $q_1$ for fixed values of $z_1$, which upon differentiation of equation (13) yield

$$(18)\quad \frac{\partial q_1}{\partial p}=\frac{p[v-C_1'-G_2-x^*(1-Z)]}{C_1''}<0.$$

This expression is negative since, as was shown in Section III.A, the unit cost of low quality $G(Z)+x^*(1-Z)$ is an increasing function of $x^*$. Since the sum of the first- and second-period terms in equation (13) is zero, and since these quality-related terms represent the only difference between the two expressions, increasing $q_1$ must have a negative effect on the marginal profits resulting in the second period. If this were not the case, the irreversibility constraint would not be binding since additional profits would be yielded by increasing output in period 2 when the penalty rose to $x^*$.

The probability $p$ of a penalty increase also has an unambiguous direct effect on $z_1$, for fixed values of $q_1$, since

$$(19)\quad \frac{\partial z_1}{\partial p}=\frac{\beta Q(-G_1'+x_*)}{\pi_{z_1z_1}}>0$$

(since $G_1'$ exceeds $x_*$, see equation (16)). Increasing the chance of an increase in the regulatory penalty dampens the initial output level and raises the initial quality investment. These results always hold if there is only one choice variable, or in situations in which there are competing effects, if the direct effects are dominant.

---

[12] For example, in the case of $\partial q_1/\partial p$ the value obtained when effects through both choice variables are considered is negative except for the presence of a term $\beta Q(-G'+x)(G'-x_*)$ generated by the indirect effect of $p$ on $q_1$ through $z_1$ that is positive since, from equation (16), we have $x>G_1'>x_*$. The underinvestment in quality increases the expected regulatory penalties on capacity investments, leading to a possible reversal in the $\partial q_1/\partial p$ result if this effect is large.

### B. *The Effect on Profits and the Value of Information*

Since regulatory uncertainty alters enterprise decisions in a manner that is quite different from what would result if penalties in each period were equated to their expected levels, profits will also be affected. If the regulatory penalty $x$ fully reflects the value of the quality component to society, the value of profits will equal the net surplus to society and can serve as an economic efficiency measure. To analyze the implications for firms' profitability, let us introduce the notation $\hat{\pi}$ to denote the present value of $\pi$ obtained over both periods when the firm selects its choice variables optimally. The analogous values in period 2 will be $\hat{\pi}_2(x^*)$ for second-period profits if $x^*$ is the penalty and $\hat{\pi}_2(x_*)$ after the penalty lottery has resulted in $x_*$.

The value of $\hat{\pi}$ depends on $p$ directly and through its influence on the choice variables $q_1(p)$ and $z_1(p)$, which in turn affects $Q(x_*, z_1)$. When differentiating $\hat{\pi}$ with respect to $p$, all terms involving $\partial q_1/\partial p$ and $\partial z_1/\partial p$ will drop out (by the envelope theorem) since the investment levels will be chosen optimally in response to changes in any parameter such as $p$. The effect of $p$ on profits is consequently given by

$$\partial\hat{\pi}/\partial p = \beta\left[\hat{\pi}_2(x^*) - \hat{\pi}_2(x_*)\right] < 0,$$

since profits are lower in the state with a higher regulatory penalty (i.e., $\hat{\pi}_2(x^*)$ is below $\hat{\pi}_2(x_*)$). Increasing the chance that the regulatory penalty will rise reduces the discounted expected profits, as one might expect. The second derivative of the profit function indicates the curvature of this relationship, where

$$\partial^2\hat{\pi}/\partial p^2 = \beta(\partial q_1/\partial p)\left[v - C_1' - G_2 - x^*(1-Z)\right] + \beta\frac{\partial_{z_1}}{\partial p}Q(G_1' - x_*) > 0,$$

which implies that $\hat{\pi}(p)$ is convex.[13]

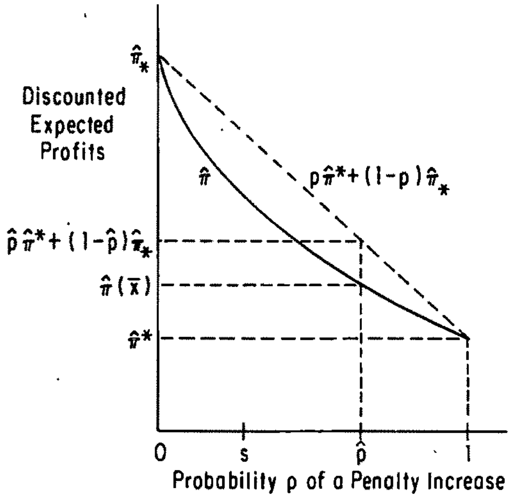


FIGURE 1. THE EFFECT OF REGULATORY UNCERTAINTY ON PROFITS

The function $\hat{\pi}$ is sketched in Figure 1. The two endpoints of the curve are the profits $\hat{\pi}_*$ with a certain drop in the penalty to $x_*$ and the profits $\hat{\pi}^*$ with a certain increase in the penalty. The weighted average of these points is the dashed line that lies above the $\hat{\pi}$ curve. From Jensen's Inequality, we have the result that $p\hat{\pi}^* + (1-p)\hat{\pi}_*$ exceeds $\hat{\pi}(\bar{x})$, where these values are indicated on the diagram for one possible value of $p$, denoted by $\hat{p}$. The firm facing an uncertain regulatory policy lottery $(p, x^*; 1-p, x_*)$ will make lower profits than if it faced the penalty $px^* + (1-p)x_*$ with certainty.

A related issue is how the size of this loss is affected by the value of $p$. In particular, what is the expected value of perfect information *EVPI* regarding the future regulatory policy or, viewed somewhat differently, what is the minimal value of the expected opportunity loss that will be incurred? The *EVPI* value is simply the gap between the two curves in Figure 1, or

$$EVPI = p\hat{\pi}^* + (1-p)\hat{\pi}_* - \hat{\pi}.$$

---

[13] The direction of this inequality can be easily verified. As noted above, the value of the derivative of second-period profits with respect to $q_1$ is negative since the irreversibility constraint is binding so the bracketed term is negative. From equation (16), $G'$ exceeds $x_*$, and the signs of $\partial q_1/\partial p$ and $\partial z_1/\partial p$ are negative and positive, respectively, assuming the direct effects of $p$ are dominant.

The gap will reach its widest amount at some interior value of $p$. The effect of $p$ on *EVPI* is given by

$$\partial EVPI/\partial p=(\hat{\pi}^{*}-\hat{\pi}_{*})-\partial\hat{\pi}/\partial p.$$

For small values of $p$, this expression is positive since the slope of the profit function $\hat{\pi}$ is steeper than the weighted average curve. The losses due to uncertainty are greatest when $\partial\hat{\pi}/\partial p$ equals $(\pi^{*}-\pi_{*})$, which is the slope of the weighted average curve, and *EVPI* decreases thereafter. The point $s$ in Figure 1 represents the $p$ value at which maximum efficiency losses occur; $s$ need only be some interior point and is not necessarily 1/2.

One would expect that a firm's profits would be decreasing with respect to $p$ wholly apart from any losses from uncertainty since higher regulatory penalties increase its costs. The more interesting result is that uncertain regulatory policies impose additional opportunity losses as firms are discouraged from making irreversible investment commitments. These losses are greatest when the lottery on regulatory penalties involves an intermediate $p$ value rather than an extreme probability close to zero or one.

## IV. Conclusion

Three separate effects of regulatory policies on enterprise decisions can be distinguished. First, increased regulatory penalties will diminish output and profits in static models or, equivalently, in multiperiod models in which investments are completely reversible, as is well known. Second, if investment commitments are irreversible, there will be an additional effect of a known schedule of changes in the regulatory policy, which will depress output even further. Finally, the addition of uncertainty with regard to regulatory policy produces a third effect resulting in expected opportunity losses for firms. Both output and quality investments will be depressed by regulatory policy lotteries. Regulations influence current enterprise decisions not only through their current level, but through their expected future level and the degree of uncertainty regarding these future regulatory policies.

## REFERENCES

**Arrow, Kenneth J.,** "Optimal Capital Policy with Irreversible Investment," in J. N. Wolfe, ed., *Value, Capital and Growth*, Chicago: Aldine Publishing Co., 1968, 1–19.

**Denison, Edward,** "The Puzzling Drop in Productivity," *Brookings Bulletin*, Fall 1978, *15*, 10–12.

**Haveman, Robert, and Christainsen, Gregory,** "Environmental Regulations and Productivity Growth," in H. Peskin et al., eds., *Environmental Regulation and the U.S. Economy*, Baltimore: Johns Hopkins University Press, 1980, 55–76.

**Kosters, Marvin,** "Government, Regulation, and Productivity," Working Paper No. 1, The American Enterprise Institute, 1980.

**MacAvoy, Paul,** *The Regulated Industries and the Economy*, New York: W. W. Norton, 1979.

**Pashigian, Peter,** "Environmental Regulation: Whose Self Interests are Being Served?," Graduate School of Business Working Paper, University of Chicago, 1981.

**Portney, Paul R.,** "The Macroeconomic Impacts of Federal Environmental Regulation," in H. Peskin et al., eds., *Environmental Regulation and the U.S. Economy*, Baltimore: Johns Hopkins University Press, 1980, 25–54.

**Schultz, George P.,** "The Abrasive Interface," in John T. Dunlop, ed., *Business and Public Policy*, Cambridge: Harvard University Press, 1980, 12–22.

**Viscusi, W. Kip,** *Risk by Choice: Regulating Health and Safety in the Workplace*, Cambridge: Harvard University Press, 1983.

# Scale and Productivity Measurement under Risk

*By* ROBERT G. CHAMBERS*

The literature on risk and uncertainty gives an important place to studies comparing behavioral and comparative static results obtained under certainty and uncertainty. The standard neoclassical theory of the firm, however, has much more to offer than behavioral results. For example, the optimizing behavior of the profit-seeking firm insures that the degree of the firm's returns to scale can be measured by the ratio of cost to revenue. Furthermore, it is also possible to isolate measures of the rate of technical change for a profit-maximizing firm which depend solely upon observed prices and quantities.

This note demonstrates some of the implications of uncertainty for measuring scale economies and the rate of technical change. Specifically, using a model developed by Agnar Sandmo (1971), it is demonstrated that under conditions of risk and uncertainty, there are generally no measures of the scale elasticity or the rate of technical change which can be derived from observations on prices and quantity without information on either the structure of production or the utility function which underlies Sandmo's analysis.

## I. Measuring Scale

The firm is assumed to maximize the expected utility of profit. The utility of profit is denoted by $u(\pi)$, where $u(\cdot)$ is a concave, continuous, and differentiable function of profit, $\pi$. Profits are given by $\pi = pF(x,t) - w \cdot x + A$, where $F(x,t)$ is the firm's production function, $x$ is a vector of variable inputs, $w$ is a vector of nonnegative input prices, $t$ is an index of time, and $A$ can be viewed as existing wealth, fixed costs, or income from other sources.

It is assumed that $p = \bar{p} + e$, where $e$ is a random variate defined over the interval $[a, b]$, $Ep = \bar{p}$, and $e$ has density $f(e)$. Under these assumptions, expected utility maximization requires

$$E\left[u'(\pi)(p\,\partial F/\partial x_i - w_i)\right] = 0, \; i = 1,2,\dots,n$$

which after a slight manipulation can be rewritten as

$$(1) \qquad \partial F/\partial x_i = w_i/(\bar{p} + \theta), \qquad i = 1,2,\dots,n,$$

where $\theta = [Eu'(\pi)e]/Eu'(\pi)$, which is negative for a risk-averse firm since it can also be written as $\theta = \mathrm{Cov}(u', e)/Eu'(\pi)$ if $e$ has zero mean.

Recognizing that the elasticity of scale, that is, $\varepsilon = \partial \ln F(\lambda x, t)/\partial \ln \lambda|_{\lambda=1}$ can be written as the sum of the logarithmic derivatives of $F$ (see C. E. Ferguson, 1971) and using (1) obtains

$$(2) \qquad \varepsilon = \sum_i w_i x_i / (\bar{p} + \theta) y.$$

Thus, to obtain a measure of the elasticity of scale from data that is consistent with expected utility maximization, one should know $\theta$, which in turn implies a high degree of knowledge about $u$. Unlike the case of profit maximization, one apparently cannot simply use observations on ($p$, $w$, $x$, and $y$) to obtain a measure of $\varepsilon$.

It is easy to rewrite (2) as

$$(3) \qquad (\bar{p} + \theta) y = \varepsilon^{-1} \sum_i w_i x_i.$$

From expression (3) we can make some inferences about the relationship between returns to scale, $\theta$, and expected variable profit

---

*Associate Professor, Department of Agricultural and Resource Economics, University of Maryland, College Park, MD 20742. This paper is Scientific Article No. A3451, Contribution No. 6524, of the Maryland Agricultural Experiment Station. Comments from T. Phipps, R. Kramer, U. Vasavada, and D. Hueth are appreciated.

$(\bar{p}y - \Sigma_i w_i x_i)$. Namely, by (3) we can write expected variable profit as

$$(4)\quad E(\pi - A) = \frac{(1-\varepsilon)}{\varepsilon}\sum_i w_i x_i - \theta y,$$

where $E$ is the expectation operator. Using the fact that $\varepsilon$ is the ratio of average variable cost to marginal cost then implies that expected variable profit is positive so long as $(1-\varepsilon) > \theta/MC$, where $MC$ is marginal cost. Thus, so long as $F(x,t)$ exhibits either constant or decreasing returns, expected variable profit for the risk-averse firm is always nonnegative. Since $\theta < 0$ for the risk-averse firm, however, the existence of scale economies (increasing returns, $\varepsilon > 1$) does not necessarily imply negative variable expected profit. Thus, it appears there are regions over which increasing returns and positive expected variable profit are consistent with one another.

## II. Measuring the Rate of Technical Change

The rate of technical change $\dot{T}$ is defined as $\dot{T} = [\partial \ln F(x,t)]/\partial t$, or as the percentage shift in the production function attributable solely to the passage of time. Charles Hulten demonstrated that, for the profit-maximizing firm, an appropriate measure of $\dot{T}$ is given by

$$(5)\qquad \dot{y} - \sum_j w_j x_j \dot{x}_j / py,$$

where dots over variables denote time derivatives. The important thing to note about (5) is that once discrete approximations to the time derivatives ($\dot{y}$ and $\dot{x}_j$) are made, one can derive measures of the rate of technical change solely from observable data on prices and quantities.

Under the assumption of expected utility maximization the appropriate measure of the rate of technical change is given by

$$(6)\quad \dot{y} - \sum_j \frac{w_j x_j}{(\bar{p}+\theta)y}\dot{x}_j = \dot{y} - \varepsilon \sum_j \frac{w_j x_j}{c}\dot{x}_j,$$

where $c = \Sigma_j w_j x_j$ and the second equality follows from (2).

From (6) it follows that one cannot obtain a measure of the rate of technical change based solely on observable data for the expected utility-maximizing firm without making an explicit assumption on either the underlying utility function or the degree of returns to scale. Also note that the second expression in (6) is equivalent to that obtained in measuring the rate of technical change for a cost-minimizing firm. Only in the case of constant returns to scale can a measure of $\dot{T}$ be based solely on observable data.

## III. The Error of Using Observable Data

It is obvious from expressions (2) and (6) that the standard measures of the scale elasticity (cost to returns) and the rate of technical change (5) are correct for data generated by the expected utility-maximizing firm with almost probability zero. This follows by the fact, for example, that (5) only equals (6) when $p = \bar{p} + \theta$, which in turn requires

$$(7)\qquad e = \theta.$$

Unfortunately, I have not been able to establish whether the traditional measures of the scale elasticity or the rate of technical change over or underestimate (2) and (6). The bias associated with using the cost-revenue ratio instead of (2) to measure $\varepsilon$ is

$$(8)\quad \frac{\sum_i w_i x_i}{y}\int_a^b \left(\frac{\theta - e}{p(\bar{p}+\theta)}\right) f(e)\, de,$$

and the bias associated with using (5) instead of (6) is

$$(9)\quad -\sum_j \frac{w_j x_j \dot{x}_j}{y}\int_a^b \left(\frac{\theta - e}{p(\bar{p}+\theta)}\right) f(e)\, de,$$

which in general do not seem to be either positive or negative. Little further can be said about (8) and (9). However, if factor usage is increasing rapidly over time ($\dot{x}_j > 1$ for all $j$), then the absolute value of bias of inappropriate scale measure is less than the absolute value of the bias of the inappropriate productivity measure.

## IV. Indirect Measures of Scale and the Rate of Technical Change

From the dual approach to microeconomics, it is well known that one can obtain measures of $\varepsilon$ and $\dot{T}$ by simple differentiation of an appropriate indirect objective function. Matters are apparently not quite that simple under uncertainty. Define the indirect expected utility function ($V(\cdot)$) by

$$V(\bar{p},w,\sigma,A,t)=\underset{x}{\operatorname{Max}}\,Eu(\pi),$$

where $\sigma$ represents the parameters of the distribution of $e$. By the envelope theorem,

$$(10)\qquad y^{*}=\frac{(\partial V/\partial \bar{p})}{(\partial V/\partial A)}=F(x^{*},t),$$

$$(11)\qquad x_i^{*}=-\frac{(\partial V/\partial w_i)}{(\partial V/\partial A)},\qquad i=1,2,\dots,n.$$

Asterisks denote the optimal elements of the solution to the expected utility-maximization program. Using (2) in conjunction with (10) and (11) yields

$$(12)\qquad \varepsilon=-\sum_j\frac{w_j}{(\bar{p}+\theta)}\left(\frac{\partial V/\partial w_j}{\partial V/\partial \bar{p}}\right).$$

Thus, even if one is able to specify an appropriate $V$, one cannot obtain a measure of the scale elasticity without a priori knowledge on $\theta$.

Using the second equality in (10) yields

$$\begin{aligned}(13)\qquad \dot{T}&=\frac{\partial \ln y^{*}}{\partial t}-\sum_j\frac{w_jx_j^{*}}{(\bar{p}+\theta)y^{*}}\frac{\partial \ln x_j^{*}}{\partial t}\\ &=\frac{\partial}{\partial t}\left(\frac{\partial V/\partial p}{\partial V/\partial A}\right)\\ &\quad-\sum_j\frac{w_j}{(\bar{p}+\theta)}\frac{\partial V/\partial w_j}{\partial V/\partial \bar{p}}\left[\frac{\partial}{\partial t}\left(\frac{\partial V/\partial w_j}{\partial V/\partial A}\right)\right]\\ &=(\bar{p}+\theta)^{-1}\left(\frac{\partial V/\partial t}{\partial V/\partial \bar{p}}\right),\end{aligned}$$

where the last equality is seen most easily by an application of the envelope theorem. Again, it follows that the rate of technical change apparently cannot be deduced solely from knowledge of $V$. The results in expressions (10) through (13) are important since they question the ability of more modern indirect approach to obtain measures of important parameters of the underlying technology or preference structure under risk.

## V. Conclusion

The results in this paper demonstrate that apparently it is not possible to simultaneously measure scale and rate of technical change from observed data (i.e., without estimation) unless one is willing to make a specific assumption about the form of either $u$ or $F$. In the case of certainty, one is always able to obtain measures of $\varepsilon$ and $\dot{T}$ that are based solely on observed data so long as there exists either constant or decreasing returns to scale. Under uncertainty, solely data-based observations on $\dot{T}$ are only available when the underlying technology exhibits constant returns. Although these results concentrate on continuous measures of technical change, similar arguments would produce entirely analogous results for the discrete productivity indexes suggested, for example, in the recent studies by W. Erwin Diewert (1976, 1978) and D. W. Caves, L. R. Christensen, and Diewert (1981) under expected utility maximization. Finally, one should note the extremely simple form of uncertainty considered in this paper. More complicated stochastic structures (for example, nonmultiplicative random production) are bound to raise further problems.

## REFERENCES

**Caves, D. W., Christensen, L. R. and Diewert, W. E.,** "A New Approach to Index Number Theory and the Measurement of Input, Output, and Productivity," Social Systems Research Institute Workshop Series No. 8112, University of Wisconsin-Madison, 1981.

**Diewert, W. Erwin,** "Exact and Superlative Index Numbers," *Journal of Econometrics*, May 1976, *4*, 115–45.

______, "Superlative Index Numbers and Consistency in Aggregation," *Econometrica*, July 1978, *46*, 883–900.

**Ferguson, C. E.,** *Neoclassical Theory of Production and Distribution*, New York: Cambridge University Press, 1971.

**Hulten, Charles R.,** "Divisia Index Numbers," *Econometrica*, November 1973, *41*, 1017–26.

**Sandmo, Agnar,** "On the Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, *61*, 65–73.

# Welfare Measurement in the Household Production Framework

*By* NANCY E. BOCKSTAEL AND KENNETH E. MCCONNELL*

The household production approach to consumer behavior, developed from the work of Gary Becker, William Gorman, and Kelvin Lancaster, has considerable descriptive appeal in modelling the decisions of households. The approach derives from the observation that households frequently purchase market goods that do not yield utility directly, but are combined to produce commodity service flows which the household values. Thus observed behavior is determined by household production technology as well as by tastes. The advantage of this distinction is that we can pose reasonable hypotheses about characteristics of technology, though we rarely possess useful a priori information regarding tastes.

The putative advantages of the household production approach are questioned on empirical and conceptual grounds by Robert Pollak and Michael Wachter (1975). They show that jointness in production or nonconstant returns to scale cause implicit commodity prices to depend on both tastes and technology, raising serious econometric difficulties in the estimation of commodity demand functions. In addition, since commodity prices become functions of the commodity bundle consumed, the analogy to traditional demand theory breaks down.

Joint production occurs when a good enters several production processes simultaneously, or, equivalently, when a good in one production process also enters directly into the individual's utility function. The most common example is time, which provides the context for all production processes and is often associated with the production of several commodities simultaneously. Since joint production in the household is likely to be pervasive, the critique by Pollak and Wachter cannot be ignored. In response to the comment by William Barnett, Pollak and Wachter (1977) suggest dispensing with the notion of commodity prices and treating the demand for commodities as a function of goods prices. This approach confounds tastes and technology, but it eliminates the troublesome concept of commodity prices as parameters when, in fact, they are likely to be endogenous.

In this paper we show that results from positive analysis, such as the critique by Pollak and Wachter, have implications for the use of the household production framework for welfare analysis. The household production function approach has had considerable appeal for measuring welfare effects of public actions in the environmental and natural resource areas (Gardner Brown, John Charbonneau, and Michael Hay; Elizabeth Wilman). Yet traditional approaches to welfare measurement are frequently inapplicable. We argue that welfare measurement in this framework is complicated by the difficulties of unravelling tastes and technology.

We extend Pollak and Wachter's results by demonstrating that Marshallian demand functions for commodities cannot be uniquely defined. Thus Marshallian functions cannot be used to derive exact compensated functions in the manner of Jerry Hausman, and of George McKenzie and I. F. Pearce, nor can compensating and equivalent variation measures be bounded by Marshallian consumer's surplus estimates following Robert Willig. In fact, duality results that normally allow us to move between Marshallian and Hicksian functions are not applicable for commodities in the household production framework.

---

*Assistant and Associate Professors, respectively, Department of Agricultural and Resource Economics, University of Maryland, College Park, MD 20742. This paper is Scientific Article No. A3404, Contribution No. 6476, of the Maryland Agricultural Experiment Station. We wish to thank Darrell Hueth, James Opaluch, V. Kerry Smith, and Elizabeth Wilman for comments on an earlier draft.

Because Marshallian demand functions for household produced commodities are not unique, we are forced to develop welfare measures from observations on purchases of goods; that is, inputs into the household's production process. We show that under appropriate conditions, changes in the area under the demand curve for goods can serve as welfare measures for changes in the quantities of public goods. This finding is related to Hajime Hori's result that when all technology is known, the demand for public goods can be inferred from the purchases of private goods. However we develop a basis for measuring the value of public goods which, unlike Hori's measure, does not require that we know a priori the household technology. All the information that is required will be embodied in the derived demand for goods. Thus we show that the goods market is the only market which provides an indirect means of valuing changes in public goods.

The following three sections provide the three principal results of the paper. In Section I, we show that traditionally conceived Marshallian demands are not uniquely defined in this approach. In Section II, we show, in contrast, that utility-constant marginal value functions that are dependent only on preferences and not technology do exist and have the usual normative interpretation. Regardless of joint production or nonconstant returns to scale, the area behind the marginal value and marginal cost curves measures economic surplus. Changes in this area measure welfare effects associated with changes in the individual's economic environment. In Section III, we show that equivalent measures of welfare change can often be obtained in the market for goods which serve as inputs into the household production process. This result is analogous to the work by Richard Just, Darrell Hueth, and Andrew Schmitz, and by Just and Hueth, who show that in competitive markets, welfare changes can be measured in the markets for inputs or outputs. Additionally, we are able to obtain these measures of welfare without assuming that technology is known. Thus we provide a new conceptual basis for welfare measurement of nonmarket goods.

## I. Positive Economics and the Household Production Function

An economic model is a useful positive tool if it implies theorems about behavior which can be tested. We argue that very little structure and a paucity of testable hypotheses emerge from the household production function. This argument is consonant with the results of Pollak and Wachter, who argue that commodity demand as a function of commodity price is not a meaningful concept. We expand their arguments by demonstrating that a unique Marshallian demand curve, as traditionally conceived, cannot be derived for commodities produced in the household production framework. In addition, we argue that commodity demands as a function of goods prices provide few qualitatively predictable results.

To characterize the household production approach, suppose that the consumer enjoys an $m$-dimensional bundle of commodities $z$ which enters the quasi-concave preference function $u(z)$. Goods, denoted $x=(x_1,\dots,x_n)$, are purchased at market prices, denoted $r=(r_1,\dots,r_n)$, and are combined to produce the commodities according to the production process $t(z,x)=0$. One of the $x$'s will typically be time and can take some function of the wage rate as its price. Joint production occurs when the technology implied by $t(z,x)$ cannot be expressed in terms of separate production functions. We do not rule out joint production nor the possibility of nonconstant returns to scale.

Positive analysis derives from the solution to the representative household's income-constrained problem

$$(1) \qquad \max u(z),$$

$$\text{subject to} \qquad C(z,r)-y=0,$$

where the joint cost function $C(z,r)=\min_x\{r\cdot x \mid t(z,x)=0\}$ and $y$ is money income. In general, the budget constraint will be nonlinear in the $z$'s, and the marginal costs of commodities will be functions of the commodity bundle chosen.

Because the Marshallian demand curve has held such a central position in comparative statics and normative analysis, there have been strenuous efforts to derive this concept from the first-order conditions of problem (1). In its traditional formulation, the Marshallian demand function relates the quantity of a commodity purchased to the price of the commodity, with money income, prices of other commodities, and all other relevant variables assumed constant. The demand curve is traced out by recording the amount of the commodity the consumer would be willing to purchase at each price.

In the household production formulation, the marginal cost of producing the commodity is analogous to price in the traditional case. The marginal cost of producing a commodity is not in general constant, however, and as such does not encode all the necessary information required to ensure that all first-order conditions, including the budget constraint, will be satisfied. Holding income and other marginal cost functions constant, we attempt to trace out a Marshallian demand curve for a commodity by altering the cost of the final unit consumed of that commodity. This approach, however, cannot uniquely define the quantity consumed unless it incorporates knowledge of the entire cost function (and thus technology) or specific assumptions about how the cost function changes with changes in marginal costs.

By way of demonstration, consider the following expression for the cost function. Suppose that when no commodities are produced, costs are zero, $C(0,r)=0$. Then there exists a $\theta$ in the unit interval such that

$$(2)\quad C(z,r)=\sum_i C_i(z,r)z_i - \sum_i\sum_j C_{ij}(z(1-\theta),r)z_iz_j/2.$$

From this relationship, the first-order conditions of problem (1) become

$$(3)\quad u_i(z)-\lambda C_i(z,r)=0\ \forall i,$$

$$y-\sum_i C_i(z,r)z_i + \sum_i\sum_j C_{ij}(z(1-\theta),r)z_iz_j/2=0,$$

where $\lambda$ is the multiplier associated with the budget constraint. Unless the production function is nonjoint and homogeneous of degree one, the $C_{ij}$ in (3) will be nonzero.

Now suppose we attempt to trace out a Marshallian demand curve by asking the question: how much will be purchased at different levels of marginal cost $(C_i)$? If the hypothetical changes in marginal cost are generated by changes in input (goods) prices or other parameters of the cost function, they will, except in special instances, alter $C_{ij}$ terms as well as other marginal cost functions through changes in the $z$ vector. Changes in these additional terms in the budget constraint will make it impossible to define uniquely, at each marginal cost, a level of demand for which the income constraint holds.

A related and equally debilitating aspect of the household production function is that few qualitative comparative statics results are generated. In general, the decision function for a choice variable is obtained by differentiating the decision maker's indirect objective function with respect to an economic parameter. In such cases, appealing to second-order conditions (specifically, the curvature properties of the utility-constant expenditure function) allows us to sign the change in the choice variable with a change in the parameter.

In the household production approach, however, the expenditure function, defined as

$$m(r,u^0)=\min_z\{C(z,r)|u(z)=u^0\},$$

fails to yield comparative static results. The joint cost function $C(z,r)$ will not in general be linearly homogeneous in $z$. Thus, there will be no parameter $\beta$ such that $\partial m/\partial\beta=z_k$ and ordinary envelope theorem derivations will not follow. As a consequence there are no a priori expectations on the sign of any $\partial z_i/\partial r_k$. Pollak and Wachter's suggestion that the demand for commodities be analyzed in terms of goods prices rather than commodity prices will yield constructs which are useful for prediction but which provide no hypotheses regarding the signs of any coefficients in commodity demand functions.

The nonlinearity of the cost function also prevents the Cournot and Engel aggregations from providing useful restrictions on demand systems. Analogous to these "adding-up" theorems, we can derive the following from the constraint $C(z,r)-y=0$ and from the fact that $C(z,r)=r\cdot x$:

$$\sum_{i=1}^{m} z_i C_i \eta_i / y = 1$$

and
$$\sum_{k=1}^{n}\sum_{i=1}^{m} z_i C_i \varepsilon_{ik} / y = -1,$$

where $\varepsilon_{ik} = \partial \log z_i / \partial \log r_k$ and $\eta_i = \partial \log z_i / \partial \log y$. Unlike the standard neoclassical case with fixed prices, the ratios $z_i C_i / y$ depend on technology, do not in general sum to one, and do not possess the useful interpretation of budget shares. Only the homogeneity restrictions $\Sigma_{k=1}^{n} \varepsilon_{ik} = -\eta_i$ are retained.

Thus two conclusions arise in the positive economics setting. First, a unique Marshallian demand curve for commodities does not exist. Second, theory fails to provide useful prior restrictions on any function which relates consumption of the $z$'s to parameters in the system.

## II. Welfare Analysis in the Commodity Market

Pollak recognized that economists may wish to "use a household production framework to analyze the harm done by air pollution or the benefits of an outdoor recreation or child health project..." (1978, p. 286). This approach, depending as it does on the distinction between purchased goods and consumed commodities, has particular appeal for measuring nonmarket benefits derived from public goods (see, for example, Hori and Wilman). However the conceptual limitations discussed above preclude the straightforward application of well-known welfare measurement techniques in the household production framework.

Work by Willig, Hausman, and others employs duality results to demonstrate that, because of the link through the expenditure and utility functions, the unobserved compensated demand function can always be derived from knowledge of the observable Marshallian function. In the household production function approach, the fact that Marshallian demands are not unique prevents us from using duality to derive the compensated demand function. It also suggests that the compensated function may be undefined.

In this section we demonstrate three results. First, we argue that unlike Marshallian demand curves, compensated demand curves which reflect households' marginal valuations of the commodity do exist, independent of the cost function for producing commodities. Second, we derive a measure of surplus associated with a commodity as the area between the marginal cost and marginal value curves in commodity space. Finally, we demonstrate that the well-established conditions for measuring the compensating variation of a change in an exogenous (for example, publicly supplied) factor are applicable in this framework.

When the household is viewed as minimizing the cost of obtaining a given utility level, independent marginal value and marginal cost functions are identifiable. Consider the utility-constrained, cost-minimization problem

$$\min_{z} \{ C(z,r) | u^0 = u(z) \}$$

which produces the first-order conditions

$$C_i(z,r) - \mu u_i(z) = 0 \forall i,$$
$$u^0 - u(z) = 0,$$

where $\mu$ is the multiplier associated with the constraint on utility. These first-order conditions could be solved for reduced-form demand functions for the $z$'s as functions of goods prices, and technological and preference parameters. They also, however, allow for the determination of independent marginal cost and value functions. To see this, we need only recognize that the above problem is equivalent to one where an imaginary market with fixed parameter prices intervenes between the production and consumption activities of the household. Since the budget constraint is not required to hold along the compensated demand curve, the system of compensated curves is not dependent on the total cost function. Consequently the circumstances which cause the Marshallian curve to be ambiguous in this framework do not directly affect the compensated demand curve.

We are prevented from deriving the Marshallian curve from the compensated curve by the absence of exogenous prices. Both the utility and expenditure functions exist, but the absence of prices prevents the use of Roy's Identity to derive the Marshallian curve from the indirect utility function. Also, it is impossible to move from a compensated demand function to a unique expenditure function because of the nonlinearity of the cost function. Several different cost functions (generated by different technologies), and thus different expenditure functions, can be associated with the same values of marginal costs.

Nonetheless, theoretical welfare measures can be derived in commodity space but in a way which differs from the traditional approach. For simplicity we focus on a measure for $z_1$, though any $z_i$ could be chosen. Partition the commodities such that $z=(z_1,\bar{z})$ where $\bar{z}=(z_2,\dots,z_n)$. Derive an expenditure function conditional on the level of $z_1$ (as though $z_1$ were temporarily fixed) such that

$$(4)\quad E(z_1,r,u^0) = \min_{\bar{z}}\{C(z_1,\bar{z},r)|u^0=u(z_1,\bar{z})\}.$$

By the envelope theorem,

$$(5)\quad \partial E(z_1,r,u^0)/\partial z_1 = -\big(\mu u_1(z_1,\bar{z}^*,u^0)-C_1(z_1,\bar{z}^*,r)\big),$$

where $\bar{z}^*=\bar{z}^*(z_1,r,u^0)$ are adjusted optimally as $z_1$, $r$, and $u^0$ change. The first term on the right is the compensated marginal value function for $z_1$.[1] The second term is the marginal cost of producing $z_1$. Since expression (5) reflects the change in expenditures necessary to maintain utility level $u^0$ as $z_1$ increases, this expression will be negative for $z_1 < z_1^*$, the optimal quantity of $z_1$. Note that when $z_1$ is adjusted optimally, expression (5) equals zero and the conditional expenditure function in (4) reduces to the traditional expenditure function, since

$$(6)\quad E\big(z_1^*(r,u^0),r,u^0\big)=m(r,u^0).$$

Using the function $E(z_1,r,u^0)$, we can compute the compensating variation associated with consuming $z_1^*$. This measure reflects the change in income which would keep the consumer at his initial utility level, a situation with no access to $z_1$, if he were subsequently given the opportunity to consume $z_1^*$ at input prices, $r$, and given technology. The measure is calculated by integrating (5) from 0 to $z_1^*$ yielding

$$(7)\quad \int_0^{z_1^*}\big[\partial E(z_1,r,u^0)/\partial z_1\big]\,dz_1 = E\big(z_1^*,r,u^0\big)-E\big(0,r,u^0\big).$$

This expression is the negative of the area between the compensated marginal value and marginal cost functions for $z_1$, where that area can be expressed as

$$(8)\quad A=\int_0^{z_1^*}\big[\mu u_1(z_1,\bar{z}^*,u^0) - C_1(z_1,\bar{z}^*,r)\big]\,dz_1.$$

Thus graphical measures of welfare exist in concept in the commodity market.

---

[1] It might be argued that one can integrate this inverse compensated demand function back to the distance function and then use the distance function to derive the Marshallian marginal value function (see, for example, Angus Deaton and John Muellbauer, ch. 2). Consider the distance function $d(z,u^0)$. Then

$$\partial d(z,u^0)/\partial z_i = MV_i(z,u^0),$$

where $MV_i$ is the normalized marginal value function for $z_i$, i.e., the proportion of income one is willing to pay for another unit of $z_i$. By letting $u^0=u(z)$, we have

$$MV_i(z,u^0)=MV_i(z,u(z)).$$

The term $MV_i$ changes with $z_j$ as

$$dMV_i(z,u(z))/dz_j = MV_{ij}(z,u^0)+MV_{iu}u_j.$$

This equation is an Antonelli decomposition for price-dependent demand equations. The right-hand side is composed of a utility constant slope and real income effect. Note, however, that in the household production framework, $u_j=\lambda C_j(z,r)$, so that the decomposition implies

$$dMV_i(z,u(z))/dz_j = MV_{ij}(z,u^0)+MV_{iu}\lambda C_j(z,r).$$

Thus, each cost function $C(z,r)$ implies a different slope for the Marshallian demand function.

The ultimate task of this section is to evaluate a change in an exogenous factor in the household production framework. In our illustration, let the exogenous factor be a publicly supplied or regulated good, $\alpha$, that affects the production or consumption of a household commodity. In keeping with standard welfare analysis, we interpret the compensation necessary to keep an individual at a given utility level after a change in $\alpha$ as a money measure of the associated welfare change.

Suppose that $\alpha$ is an environmental good, such as the water quality of a lake. Then it will enter the utility function directly and be complementary with some commodity which we shall denote $z_1$, such as lake recreation. In this case, utility will be a function of $\alpha$ and $z$. Alternatively, if $\alpha$ is a public health project or child care facility, it may be more appropriately viewed as an input into the production of some $z_1$, such as family health or child quality. Now $\alpha$ will enter the transformation function (i.e., $t(z,x)$ becomes $t(z,x,\alpha)$). In either case, the household's expenditure function will depend on $\alpha$.

When $u^0$ is the initial welfare level, the compensating variation of a change in the parameter vector from $\alpha^0$ to $\alpha'$ is given by

$$(9)\quad CV = m(r,u^0,\alpha') - m(r,u^0,\alpha^0).$$

Compensating variation is negative for an increase in $\alpha$ and positive for a decrease, when $\alpha$ is a desirable public good. Since the measure in (9) is not directly observable, we seek a means of evaluating the welfare effects of the changes in $\alpha$ from information on the production and consumption of the associated commodity $z_1$. The equivalence between compensating variation measures of $z_1^*$ and areas in commodity space suggests a useful approach. From the results above, we know that the difference in the areas between the marginal value and marginal cost functions evaluated for $\alpha^0$ and $\alpha'$ will be equivalent to

$$(10)\quad \int_0^{z_1^*(\alpha')}\left[\partial E(z_1,r,u^0,\alpha')/\partial z_1\right]dz_1 - \int_0^{z_1^*(\alpha^0)}\left[\partial E(z_1,r,u^0,\alpha^0)/\partial z_1\right]dz_1.$$

Expression (10) can be written as

$$(11)\quad E(z_1^*(\alpha'),r,u^0,\alpha') - E(0,r,u^0,\alpha') - E(z_1^*(\alpha^0),r,u^0,\alpha^0) + E(0,r,u^0,\alpha^0).$$

However, analogous to equation (6), the conditional expenditure function $E(\cdot)$ evaluated at the optimal value of $z_1$ is equivalent to the usual expenditure function $m(\cdot)$, so that (11) becomes

$$(12)\quad m(r,u^0,\alpha') - m(r,u^0,\alpha^0) - E(0,r,u^0,\alpha') + E(0,r,u^0,\alpha^0).$$

Thus when

$$(13)\quad E(0,r,u^0,\alpha') = E(0,r,u^0,\alpha^0),$$

the compensating variation associated with a change in $\alpha$ can be measured by changes in the area behind compensated demand and marginal cost curves for $z_1$. That is, expression (12) which equals this area collapses to the correct measure of welfare change given by expression (9).

When $\alpha$ enters the household's preference function directly, a sufficient condition for (13) to hold is that $\alpha$ be weakly complementary to $z_1$. Karl-Göran Mäler has defined weak complementary as follows: "If the demand for a private good is zero, then the demand for some environmental service [public good] will also be zero" (p. 183). Thus weak complementarity is consistent with the condition $\partial u(0,\bar{z},\alpha)/\partial\alpha = 0$, which implies that the individual is indifferent to varying levels of the exogenous good when he does not consume $z_1$. Alternatively, when $\alpha$ serves as an input into the production process, condition (13) will hold automatically if $\alpha$ is only an input in the production of $z_1$.

This section demonstrates that the compensated marginal value and marginal cost functions generated by the household production approach have normative significance and can be used to capture the welfare effects resulting from a change in exogenous factors affecting either tastes or technology. However, these results are of minimal value if there is no means of observing utility-constant marginal value functions. The usual approximation by means of consumer's surplus as well as the possibility of deriving exact measures of compensating variation from the Marshallian demand curve are precluded as well since that curve is not uniquely defined.

### III. Welfare Analysis in the Goods Market

While the commodity space of the household production framework provides conceptually valid welfare measures, our inability to observe Marshallian approximations of these measures makes normative analysis difficult. In this section we use goods space to derive equivalent and conceptually valid, but empirically feasible, measures of welfare change.

Recall that the expenditure function can be derived as

$$m(r,u^0,\alpha)=\min_x\{r\cdot x|u^0=u(z,\alpha)\}$$

where $z$ for any $x$ vector satisfies $t(z,x,\alpha)=0$ and $m(r,u^0,\alpha')-m(r,u^0,\alpha^0)$ is a money measure of the welfare effects of a change in $\alpha$. Since expenditures are linear in $x$'s, the compensated demand for some input $x_1$ is the derivative of the expenditure function with respect to $r_1$:

$$\partial m(r,u^0,\alpha)/\partial r_1=x_1(r,u^0,\alpha).$$

Define $\tilde{r}_1(\bar{r},u^0,\alpha)$, where $\bar{r}=(r_2,\dots,r_n)$, as the price that induces zero-compensated demand for $x_1$; that is, $x_1(\tilde{r}_1,\bar{r},u^0,\alpha)=0$. Note that $\tilde{r}_1$ depends on the utility level $u^0$, the level of the public good $\alpha$, and prices $\bar{r}$, although these arguments will be supressed for simplicity. The area under the compensated demand curve for $x_1$ is therefore

$$(14)\quad A=\int_{r_1^0}^{\tilde{r}_1}x_1(r,u^0,\alpha)\,dr_1,$$

$$=m(\tilde{r}_1,\bar{r},u^0,\alpha)-m(r_1^0,\bar{r},u^0,\alpha),$$

where $(r_1^0,\bar{r})$ is the prevailing price vector.

By definition, compensating variation for a change in $\alpha$ is given by expression (9), but this expression is not directly observable. Our task is to demonstrate the conditions under which expression (9) can be derived from the area in the goods market given by (14), an area which if not observable can be approximated from its Marshallian counterpart. The change in the area behind the compensated demand function for some $x_1$ caused by a change in $\alpha$ from $\alpha^0$ to $\alpha'$ can be expressed as

$$(15)\quad \Delta A=\int_{r_1^0}^{\tilde{r}_1}x_1(r,u^0,\alpha')\,dr_1$$

$$-\int_{r_1^0}^{\tilde{r}_1}x_1(r,u^0,\alpha^0)\,dr_1.$$

If

$$(16)\quad m(\tilde{r}_1,\bar{r},u^0,\alpha')=m(\tilde{r}_1,\bar{r},u^0,\alpha),$$

then $\Delta A=m(r,u^0,\alpha')-m(r,u^0,\alpha^0)$.

Thus when (16) holds, the welfare effect of a change in $\alpha$ can be measured exactly in the goods market.

Expression (16) requires that the individual be indifferent among different levels of the public good when $x_1$ is not purchased. Sufficient conditions for (16) to hold are that

(*i*) $\alpha$ is complementary to a subset of $z$ denoted $z_A$ such that $\partial u/\partial\alpha=0$ if $z_i=0$, for all $i$ in $A$;

(*ii*) $x_1$ is an essential input in the production of all $z_i$, for all $i$ in $A$. Writing the transformation as a generalized production function for $z_i$ implies $z_i=t^*(\bar{z},x)$, where $\bar{z}$ is the vector of all other $z$'s. The essentiality of $x_1$ implies $0=t^*(\bar{z},0,x_2,\dots,x_n)$ for all $\bar{z}$.

The intuition of condition (*i*) is that when the price vector induces no units of $x_1$ to be purchased, $z_i$ cannot be produced, and with no $z_i$, the individual is indifferent among different levels of $\alpha$. If $\alpha$ enters the preference function, condition (*i*) implies weak complementarity between $\alpha$ and $z_i$. If, instead, $\alpha$ enters the production function only, condition (*i*) holds trivially.

In addition, $x_1$ must be an essential input into the production of the $z_i$. Expression (16) can hold even when $x_1$ is nonessential in the production of commodities which are unrelated to $\alpha$. However the measure is incomplete if $\alpha$ is related to commodities, either through production or consumption, for which $x_1$ is nonessential. Consequently if we can conceptually measure the change in some commodity space, we can measure it in goods space as well, as long as an essential input to that commodity can be identified.

Welfare measurement in the household production framework gives further insight into the concept of weak complementarity which has played a crucial role in measuring the demand for public goods from market data. V. Kerry Smith argues that since weak complementarity describes a link between arguments of the preference function, it must be assumed and cannot be tested. He observes that behavior that appears to be consistent with weak complementarity between a public and a private good may instead be the result of technical links between these goods. However, in the household production framework, Smith's distinction between links of technology and tastes is unnecessary. Both technology and tastes affect the household's decisions regarding $z$ and thus both affect demand for goods as inputs into the production of $z$. Conditions (*i*) and (*ii*) demonstrate that it does not matter whether the link between the public good and the produced commodity is through the preference or production function, as long as the public good is of no value when the commodity is not produced.

This approach represents an advancement in the art of valuing public goods in the household production framework on two counts. First, by focusing on goods rather than commodities, it avoids the ill-defined Marshallian commodity demands. Second, it does not require that technology be known as Hori's approach does. All information about technology necessary to derive the value of a public input is embodied in the derived demand functions for goods.

This section demonstrates that when certain conditions are met, the welfare changes resulting from the change in an exogenous variable can be measured as areas behind Hicksian demand curves for goods. This result is an important corollary to the general results of Just and Hueth, and of Just, Hueth, and Schmitz, who demonstrate the duality of surpluses in factor and product markets for competitive firms. The equivalence of welfare measures in alternative markets depends on the input or output in question being "necessary." That is, when prices are such that the input is not hired or the output is not produced, the firm must be assumed to shut down. We show that parallel conditions on the essentiality of goods as inputs in the household technology allow us to value changes in public goods in the household production framework.

## IV. Conclusion

This paper develops an approach for welfare measurement in the household production function framework. While the household production function offers almost no testable hypotheses involving the production and consumption of commodities, welfare measures in commodity space do exist. However because the Marshallian demand curve is not well defined, it is not possible to estimate this demand curve. As a consequence, the practice of approximating welfare changes as the area under Marshallian demand curves is precluded. Hence, welfare measurement using commodity demand functions is not feasible in the household production framework.

The use of purchased goods in the production of commodities provides an opportunity for measuring welfare changes. When a good is essential in the production of a commodity, and a publicly controlled resource is complementary to the commodity, changes in the area under the Hicksian demand curve for that good can be interpreted as welfare measures of changes in the public resource. Thus an alternative and feasible means of measuring welfare change is provided for the household production function framework.

## REFERENCES

**Barnett, William A.,** "Pollak and Wachter on the Household Production Function Approach," *Journal of Political Economy*, October 1977, *85*, 1073–82.

**Becker, Gary S.,** "A Theory of the Allocation of Time," *Economic Journal*, September 1965, *75*, 493–517.

**Brown, Gardner, Jr., Charbonneau, John and Hay, Michael,** "The Value of Wildlife Estimated by the Hedonic Approach," Working Paper No. 6, U. S. Department of Interior, Fish and Wildlife Service, 1978.

**Deaton, Angus and Muellbauer, John,** *Economics and Consumer Behavior*, Cambridge: Cambridge University Press, 1980.

**Gorman, W. H.,** "A Possible Preference for Analysing Quality Differentials in the Egg Market," mimeo., 1956; reissued as Discussion Paper No. B4, London School of Economics Econometrics Program, 1976.

**Hausman, Jerry A.,** "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, September 1981, *71*, 662–76.

**Hori, Hajime,** "Revealed Preference for Public Goods," *American Economic Review*, December 1975, *65*, 978–91.

**Just, Richard and Hueth, Darrell,** "Welfare Measures in a Multimarket Framework," *American Economic Review*, December 1979, *69*, 947–54.

**______, ______, and Schmitz, Andrew,** *Applied Welfare Economics and Public Policy*, Englewood Cliffs: Prentice-Hall, 1982.

**Lancaster, Kelvin, J.,** "A New Approach to Consumer Theory," *Journal of Political Economy*, April 1966, *74*, 132–57.

**McKenzie, George W. and Pearce, I. F.,** "Welfare Measurement—A Synthesis," *American Economic Review*, September 1982, *72*, 669–82.

**Mäler, Karl-Göran,** *Environmental Economics: A Theoretical Inquiry*, Baltimore: John Hopkins Press, 1974.

**Pollak, Robert A.,** "Welfare Evaluation and the Cost-of-Living Index in the Household Production Model," *American Economic Review*, June 1978, *68*, 285–99.

**______ and Wachter, Michael,** "The Relevance of the Household Production Function and Its Implications for the Allocation of Time," *Journal of Political Economy*, April 1975, *83*, 255–77.

**______ and ______,** "Reply: Pollak and Wachter on the Household Production Function Approach," *Journal of Political Economy*, October 1977, *85*, 1083–86.

**Smith, V. Kerry,** "Introduction to Advances in Applied Microeconomics and Some Perspectives on Volume I," in his *Advances in Applied Micro-Economics*, Vol. 1, Greenwich: JAI Press, 1981.

**Willig, Robert D.,** "Consumer's Surplus Without Apology," *American Economic Review*, September 1976, *66*, 589–97.

**Wilman, Elizabeth, A.,** "Hedonic Prices and Beach Recreational Values," in V. K. Smith, ed., *Advances in Applied Micro-Economics*, Vol. 1, Greenwich: JAI Press, 1981.

# Inventories and Sticky Prices: Note

*By* DAVID P. SCHUTTE*

In a recent contribution to this *Review*, Alan Blinder provides an interesting analysis of the relationship between optimal inventory decisions and the process of price adjustment. Arguing that inventory permits separation of production and sale decisions, and allows the firm greater flexibility in responding to anticipated demand disturbances, Blinder shows that the responsiveness of relative prices depends critically upon the cost of storage and the temporal structure of the demand shocks.

Few economists would fault the intuitive substance of Blinder's basic proposition, and the proposition is justified with characteristic precision. The analysis, however, is considerably less general than Blinder maintains. The formal argument of his paper relies upon a simple model of a monopolist capable of backlogging orders and storing unsold goods. As the model is constructed, order backlogs assume a role of fundamental importance in any stationary equilibrium. This feature of the model vitiates Blinder's extension of his result to situations in which order backlogs are prohibited.

The basic monopoly model developed in Sections I and II of Blinder's paper places no restriction upon the sign of inventory. Negative inventory is interpreted as net unfilled orders, while no distinction is drawn between reductions in gross inventory and additions to the order backlog. As a result, the cost of holding inventory is symmetric around an arbitrary minimum point. The only role of inventory is to permit the time paths of production and sales to differ: gross inventory allows production to antedate sales, while order backlogs generate revenue before production costs are borne. Under these circumstances, if future costs and revenues are discounted, it is natural to expect any stationary equilibrium level of net inventory to be negative. This can be shown to be true of Blinder's model.

It is useful to distinguish in general between the nonstochastic stationary level of inventory ($\bar{n}$) and the stochastic stationary level ($n$). The former is realized when the current expected demand disturbance ($\varepsilon_0$) is zero and production equals sales; the latter, which differs from customary usage, represents the level of inventory expected to prevail when adjustment to the current (nonzero) expected demand disturbance is complete. If the demand disturbance is not a random walk, so that $\rho < 1$ in Blinder's (5), the two equilibria coincide. Relatively little attention is devoted to either stationary solution, because Blinder's main interest is in the optimal current solution level ($n_1$). The stochastic stationary level is crucially important, however, since it represents the goal toward which the current choice must provide an optimal first step.

To verify that both $\bar{n}$ and $n$ are negative whenever production is strictly positive, note that Blinder's necessary condition (7) with the coefficient $c_1$ set equal to zero, together with the definition of $\lambda_t$ as the deviation of the shadow value of inventory ($q_t$) from its nonstochastic stationary value ($\bar{q}$), implies

$$(1) \qquad y_t = cq_t = c(\lambda_t + \bar{q}) > 0,$$

so that

$$(2) \qquad \lambda_t > -\bar{q},$$

for all $t$ such that $y_t > 0$. From Blinder's condition (12), a stationary solution requires

$$(3) \qquad \lambda = -b(n - \bar{n})/r,$$

which, together with the preceding results, implies

$$(4) \qquad n < \bar{n} + r\bar{q}/b = 0.$$

The equality above represents Blinder's necessary condition (9), with the coefficient $b_1$ set equal to zero, evaluated at the nonstochastic stationary state. That condition itself implies

---

*Department of Economics, Texas A&M University, College Station, TX 77843.

$$\bar{n} = -r\bar{q}/b < 0, \tag{5}$$

if $\bar{y} > 0$. The stationary solutions of Blinder's model, in both the stochastic and nonstochastic versions, entail negative equilibrium levels of net inventory.

With respect to the stationary solutions, Blinder's model is better viewed not as a model of optimal inventories in the usual sense, but as a model of optimal order backlogs. This interpretation avoids the difficulty acknowledged by Blinder in Section III.B, where the "desired" level of inventory is defined in (28) to be

$$n_0^* \equiv \bar{n} - \frac{d(1+r-\rho)}{(z_2-\rho)(1-z_1)}\varepsilon_0. \tag{6}$$

Since the coefficient on the current expected demand disturbance ($\varepsilon_0$) can be shown to be negative, this definition seems paradoxically to associate positive demand shocks with lower desired inventory levels. Blinder suggests that the first-order autoregressive specification of the demand shock might be unrealistic. Recognizing that $\bar{n}$ is negative, a more natural resolution is clearly available: $n_0^*$ is to be interpreted as the desired level of *net* inventory, which is reduced by positive expected demand shocks as the desired level of gross order backlogs increases. The need for this interpretation is symptomatic of the importance attached to order backlogs in the model.

In Section V.B, Blinder claims that his analysis encompasses the asymmetric price responses associated with "stockouts" when inventory is constrained to be nonnegative. If the firm does not stock out, that is, if the optimal choice of $n_1$ is nonnegative, "...the dynamic system follows the same equations as before" (p. 346). In view of the preceding discussion, it is clear that this solution strategy is inapplicable: the stationary solution of the original system is inconsistent with the nonnegativity constraint, regardless of the implied value of $n_1$.

The requirement that inventory be nonnegative introduces the possibility of a corner solution, in which the firm would like to sell more than its fixed stock permits. The probability of this event depends upon the level of inventory chosen. Part of the expected return from holding inventory reflects the contingency of stockout, so this contingency must appear in any rational decision rule. This problem cannot be resolved simply by adding an arbitrary positive constant ($K$) to the solution for inventory, as Blinder suggests, unless the range of the demand shock is also arbitrarily restricted. The complication did not arise in Blinder's original specification precisely because negative inventory was permitted and any distinction between gross inventory and order backlogs was eschewed.

It must be emphasized that the preceding remarks do not compromise the value of Blinder's central theorem applied to the particular monopoly situation with order backlogs from which it was derived. The force of the result, however, stems partly from the very special structure of the situation under study, and extension to more general inventory management problems is far from trivial.

## REFERENCES

**Blinder, Alan,** "Inventories and Sticky Prices: More on the Microfoundations of Macroeconomics," *American Economic Review*, June 1982, *72*, 334–48.

# Intrafirm Productivity: Comment

*By* M. SHAHID ALAM*

In a recent paper in this *Review*, Harvey Leibenstein argues that "once we look at productivity from the viewpoint of X-efficiency theory postulates, effort discretion is implied, and productivity becomes a game theoretic problem" (1982, p. 92). In particular, he claims to show that "rational" choice on the part of firm members leads to a prisoners' dilemma (*PD*) outcome, a solution which is generally "held in abeyance by conventions, institutions, and laws, involving trust, enforcement of contracts, etc." (p. 96).

While all this constitutes a fresh start, it is my contention that the analysis does not stand up to critical scrutiny. First, I maintain that Leibenstein's characterization of intrafirm productivity as a *PD* problem cannot be sustained: in fact, the dilemma does not exist. At a more fundamental level, however, I find myself at variance with Leibenstein's game theoretic conception of the employment relationship.[1] Instead, I introduce the concept of an effort supply function to describe the relationship between the employees' mean effort level and the firm's offer of wage-working conditions. This leads to an alternative view of intrafirm productivity that is essentially choice theoretic. However, before I develop the critique and reformulation, a brief review of Leibenstein's model will be necessary.

## I. Production as a Prisoners' Dilemma Problem

On the assumption that all employees make the same choice, Leibenstein looks at productivity as a nonzero-sum game with two players: employees and the firm. The options of employees, consisting of effort choices, range from $\pi_1$ to $\pi_n$ in Figure 1: $\pi_1 > \pi_2 > \ldots > \pi_n$.[2] Similarly, the firm chooses from a set of wage-working conditions, ranging from $W_1$ to $W_n$: $W_1 > W_2 > \ldots > W_n$.[3] The payoffs to these different options are defined by Leibenstein as follows: first, "a shift along any of the horizontal option sets (say along $AF_1$) involves adversarial movements for which the firm gains at the expense of employees. Similarly, vertical movements from top to bottom are adversarial movements (say along $AE_1$) for which the employees gain at the expense of the firm" (pp. 95–96). It is also the case that "joint choices which are adversarial lead to smaller joint payoffs than would otherwise be the case" (p. 96).

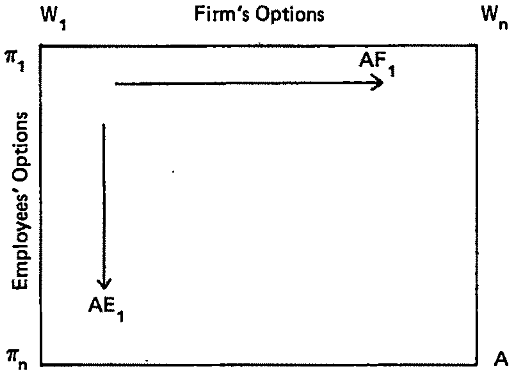


FIGURE 1

On the assumption that each side behaves "rationally" and maximizes "given the other's behavior" (p. 93), a *PD* outcome is inevitable in the above framework. If the employees choose $\pi_1$, then it certainly pays for the firm to choose $W_n$. Similarly, if the firm believes that employees will choose $\pi_n$, it still makes sense for the firm to choose $W_n$. The same logic applies to the employees. The upshot of all this is clear: rational choice leads to $W_n$, $\pi_n$, that is, the *PD* outcome.

---

*Senior Research Economist, Applied Economics Research Centre, University of Karachi, Karachi-32, Pakistan.

[1] It will be shown that the game theoretic conception results from a rather limited interpretation of "rational" choice.

[2] The level of effort represented by $\pi_1$ is forthcoming when, as Leibenstein states, "every employee acts in the best interest of the firm. He treats the firm as he would like the firm to treat him, and puts forth effort as if the enterprise was his own" (p. 92). On the other hand, $\pi_n$ is forthcoming when "the individual works as little as possible in the firm's interest and does other things (on the job) to pursue his own private interests" (pp. 92–93).

[3] Under $W_1$, the firm provides maximal benefits consistent with "sustainable profits"; under $W_n$, it attempts to minimize working conditions.

## II. Is a Prisoners' Dilemma Outcome Inevitable?

A closer look at the above characterization of intrafirm productivity reveals several difficulties. First, it may be noted that the employment relationship does not meet a necessary condition of the *PD* game, namely, that negotiation between the parties is prevented by its cost or some other circumstance. Since the firm and employees are free to negotiate, and at costs that are negligible, it appears that they will agree to some combination of payoffs which leaves both better off vis-à-vis the *PD* outcome.

Further, even if we rule out negotiations between the firm and employees, as Leibenstein does, a *PD* outcome need not follow. Adversarial movements, such as the one along $AF_1$ in Figure 1, are expected by Leibenstein, and correctly so, to lower the firm's productivity.[4] What this implies is that the posited increase in the firm's payoff as it moves along $AF_1$ takes place out of a declining total product. It appears more reasonable to suppose that the firm stands a better chance of increasing its payoff out of an increasing total product, that is, by moving (from right to left) along $AF_1$. If this were the case, the firms may choose $W_1$ (where the total product would be at its maximum), and the employees would continue to choose $\pi_n$. At $\pi_n, W_1$, the employees will enjoy their highest possible payoff and the firm too escapes its worst payoff given at $\pi_n, W_n$.

Finally, it may be shown that Leibenstein's game theoretic approach to the problem of intrafirm productivity stems from a definition of rational behavior (see his p. 93)—the maximization of one's own payoff given the other's behavior—which is a bit too narrow. Employees are neither myopic nor one-period maximizers, and, given the firm's offer of wage-working conditions, they are unlikely to respond by choosing to work "as little as possible in the firm's interest." In a more realistic model, employees would also take into account the expected loss from being fired, and, since this will depend, among other things, on the firm's wage-working conditions, the employee's effort choice will be functionally related to $W$. This functional dependence naturally imparts a choice theoretic complexion to the determination of productivity in the firm. It is this approach which I shall explore in the next section.

## III. An Alternative View

Before I develop my choice theoretic approach to the determination of intrafirm productivity, it seems appropriate to place it in perspective. An impressive body of literature has grown in the recent past which departs from the conventional neoclassical analysis of the firm as a fully specified production process "governed by the exchange of property claims in inputs and outputs" (see Herbert Gintis and Samuel Bowles, p. 342). Instead, it focuses on the characteristics of the employment relationship, and either derives its rationale,[5] or works out its implications for the effort choice of employees. The X-efficiency literature concentrates on the latter issue. In Leibenstein's words, it views "employment contracts...incomplete since remuneration is usually well specified but effort is not. Agents (employees), in principal-agent relations, need not behave exactly as the principals wish. As a consequence some effort discretion exists" (1982, p. 92). Also taking the same approach, though from a different perspective, is the work of Gintis (1976), and that by Gintis and Bowles, who also maintain that since the employment contract is incomplete and "the payer of the wage has extremely limited disposition over the worker...other than the power to terminate the labor contract..., [the cost of work] depends jointly upon the price of labor power and the capitalist's costs of extracting labor from the worker once hired" (p. 342). My exercise may be construed as an extension of this literature.

---

[4] According to Leibenstein: "it seems reasonable to presume that adversarial behavior between employees and the firm will usually decrease productivity, while cooperative behavior will increase it" (p. 94).

[5] The pioneering work in this area is of course that of Ronald Coase (1937) who perceived the rationale of the employment relationship in the reduction of transaction costs. In the same tradition, Armen Alchian and Harold Demsetz (1972) offered to explain the structure of the classical firm as a solution to the problem of shirking associated with team production. Similarly, Oliver Williamson, Michael Wachter, and Jeffrey Harris (1975) rationalize the phenomenon of the internal labor market as a response to the problems created by task idiosyncracies. More recently, Benjamin Klein has shown that a "particular form of transaction cost based upon the existence of incomplete contracts (due to uncertainty and measurement costs)—a transaction cost I have called the hold-up problem—may be an important reason in many cases for termination-at-will and exclusive-dealing contract arrangements" (1980, p. 362).

My alternative theory of intrafirm productivity may now be set out. Given the nature of the employment relationship and the firm's offer of wage-working conditions, $W$, the employee is faced with the task of choosing an effort level that maximizes his welfare. This welfare-maximization exercise has to take into account, among other things, the firm's monitoring activities and firing policy. As a part of this exercise, the employee will form an estimate of the probability, $p$, of being fired associated with different effort levels: $p$ will vary inversely with effort level. It follows that if $W^e$ is his expected income on being fired, the employee's expected loss from being fired, as pointed out by Gintis and Bowles, is $p(W-W^e)$: $W$ being greater than $W^e$.[6] It may now be seen that an increase in $W$ increases the expected loss from being fired, both directly, and, indirectly, through an increase in $p$.[7] Now if the employee was already in an effort equilibrium before the increase in $W$, he may be expected to respond to this change by increasing his effort level, thus lowering $p$ and his expected loss. In other words, we may expect the employee's actual supply of effort to be a positive function of $W$.[8]

The firm's effort supply function ($ESF$) is simply the average supply of effort by all employees at different levels of $W$. On the assumption that there is a one-to-one correspondence between effort supply and productivity (value-added) per employee, the $ESF$ may alternatively be represented as a relationship between *average* productivity per worker, $P$ and $W$. Deducting monitoring costs per employee $m$ from $P$ gives us the net effort supply function ($NESF$) facing the firm. It is with respect to this $NESF$ that the firm will choose $W$ to maximize its profits. Thus, wage-working conditions are determined endogenously in our model. This may be easily demonstrated.

In Figure 2, let the origin represent some initial wage-working conditions, $W_0$, and net productivity, $\rho(=P-m)$, corresponding to which the firm breaks even: $m$ is assumed constant throughout my analysis.[9] A 45° line is drawn through the origin. Consider several possible functions: $NESF^1$, $NESF^2$, and $NESF^3$. Given $NESF^1$, the firm will remain at the origin since any increase in $W$ is not compensated by an equivalent increase in net productivity. With $NESF^2$, however, an increase in $W$ is more than compensated by an increase in $\rho$ and the firm's final equilibrium is given at $W^2$ where its profit, equal to the vertical distance between $NESF^2$ and the 45° line, is at its maximum. With $NESF^3$, the firm may either remain at the origin if its knowledge of the relevant $NESF$ is confined to the neighborhood of its existing position;

---

[6] If $W^e$ were greater than $W$, the employee would quit his current job.

[7] The employee's subjective probability of being fired may be given by $p=p(\pi^e-\pi, m)$, where $\pi^e$ is the employee's estimate of the effort the firm expects of him, $\pi$ is the effort he supplies, and $m$ is the firm's monitoring cost: $p_1>0, p_2>0$. Further, it may be postulated that $\pi^e=\pi^e(W)$ and $\pi_1^e>0$. It follows that $\partial p/\partial W>0$.

[8] Strictly speaking, a utility analysis of this problem would be necessary, and I suspect that the response of actual effort to changes in $W$ may be ambiguous. However, this is not likely to affect the drift of my analysis. For one thing, the firm may seek to offer enhanced rewards only to employees who respond positively to them; in this way, the response of the firm's average effort to rising $W$ would be positive even though individual employees may reduce their effort in response to increases in $W$.

[9] Alternatively, one could assume that optimal levels of $m$ vary with $W$, being determined by the rule that $m$ is chosen to equate the marginal cost of monitoring with the corresponding marginal increment in productivity.

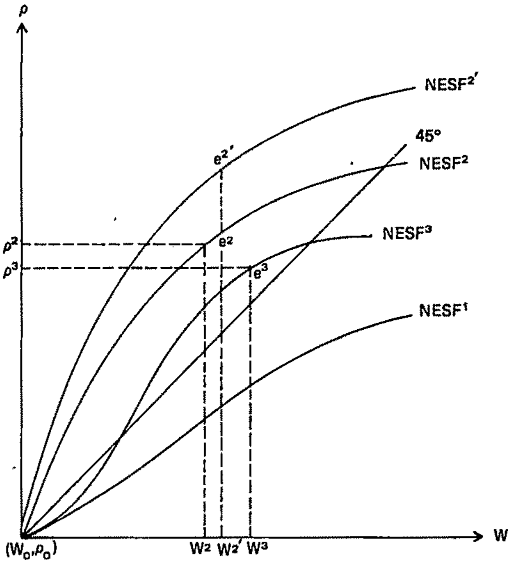


FIGURE 2

or it may attain equilibrium at $W^3$ if its knowledge of $NESF^3$ is complete.

The major conclusion to be derived from the preceding analysis is that productivity and wage-working conditions will vary across firms in a given industry even when the technology employed is identical across firms.[10] Since these differentials—defined by Leibenstein as X-efficiency—are explained in terms of neoclassical assumptions, it appears that we need not fall back upon other forms of behavior (for example, selective rationality and inert areas; see Leibenstein, 1976) to explain their presence. Some other implications of the above exercise may also be noted. It may be seen that productivity is not necessarily positively correlated with wage-working conditions: $W^3(>W^2)$ is associated with $\rho^3(<\rho^2)$. It may also be seen how a change in management can affect productivity. Given $NESF^3$, if the firm is caught in an equilibrium at the origin because of management myopia, a better-informed management could increase productivity and profits by moving to the new equilibrium at $e^3$.

Some interesting comparative static results may also be derived from the above framework. Thus, an increase in output price may raise wages. In Figure 2, an increase in price shifts $NESF^2$ proportionately to $NESF^{2'}$ and, correspondingly, the equilibrium $W$ shifts from $W^2$ to $W^{2'}$. However, if the firm is at the origin, an increase in wage may or may not be observed: the actual outcome will depend upon the position of $NESF$ with respect to the 45° line and the magnitude of the price shift.

It may also be shown that an increase in unemployment may increase productivity. Higher unemployment, it may be noted, will increase the employee's probability of being fired and also lower his expected income if fired, thereby increasing his expected loss from being fired. The employees are likely to respond to this by raising their effort level in a bid to lower their probability of being fired. Thus, productivity may increase in periods of unemployment: a phenomenon which has been observed during depressions.

## IV. Concluding Remarks

This comment develops some of the implications of the firm as a contractual relationship between firm and employees, in which the former employs its offer of wage-working conditions, monitoring activities and firing policy to optimize its extraction of effort from its employees who are likewise engaged in choosing an effort level that maximizes their welfare. However, unlike Leibenstein (1982) who models this relationship in a game theoretic mode, I have found it more useful to consider it in a choice theoretic framework. Employing the standard assumptions of rational behavior, my alternative formulation derives an effort supply function with respect to which each firm chooses its optimal wage-working conditions, which in turn determines the level of productivity. Thus productivity, rather than being technologically determined, is shown to depend on the effort choices of employees and the organization of the firm.

---

[10] Differences in $NESF$s across firms with identical technology are likely to arise from two sources: differences in the utility functions of employees and variations in managerial abilities of firms.

## REFERENCES

**Alchian, Armen and Demsetz, Harold,** "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

**Coase, Ronald H.,** "The Nature of the Firm," *Economica*, November 1937, *4*, 386–405.

**Gintis, Herbert,** "The Nature of Labor Exchange and the Theory of Capitalist Production," *Review of Radical Political Economics*, Summer 1976, *8*, 36–54.

______ **and Bowles, Samuel,** "The Welfare State and Long-Term Economic Growth: Marxian, Neoclassical, and Keynesian Approaches," *American Economic Review Proceedings*, May 1982, *72*, 341–45.

**Klein, Benjamin,** "Transaction Cost Determinants of 'Unfair' Contractual Arrangements," *American Economic Review Proceedings*, May 1980, *70*, 356–62.

**Leibenstein, Harvey,** *Beyond Economic Man: A New Foundation for Microeconomics*, Cambridge: Harvard University Press, 1976.

______, "The Prisoners' Dilemma in the Invisible Hand: An Analysis of Intrafirm Productivity," *American Economic Review Proceedings*, May 1982, *72*, 92–97.

**Williamson, O. E., Wachter, M. L. and Harris, J. E.,** "Understanding the Employment Relation," *Bell Journal of Economics*, Spring 1975, *6*, 250–80.

# Intrafirm Productivity: Reply

*By* HARVEY LEIBENSTEIN*

In my 1982 paper, an attempt is made to study the simultaneous interactions of individuals and groups within the firm in order to show that, if effort discretion exists, then the productivity problem could be viewed as an intrafirm game that will usually have a suboptimal solution. Since the paper was written for the AEA annual meetings, space constraints necessitated forcing the ideas into a terse mode. This may have been responsible for some misunderstandings. In this reply, I will try to clear up some basic misinterpretations, and then touch on, what appears to me to be, some difficulties with the effort supply concept.

In his comment, M. Shahid Alam implies that I believe the prisoner's dilemma outcome is inevitable. In fact, I argue the opposite. The ideas of effort discretion, prisoners' dilemmas, and effort conventions have to be taken simultaneously. Section I of my paper suggests that a behind-the-scenes *latent PD* problem exists. Section II argues that *effort conventions* are substitutes for the latent *PD* outcome.

A basic difference between Alam and myself is his belief that the prisoner's dilemma problem depends on a lack of communication. A lack of communication may be a sufficient condition for a prisoner's dilemma to exist given certain characteristics of the payoff matrix, but it is not a necessary one. Equally important is a lack of trust so that enforceable agreements may not be readily achievable. The mere fact that the parties can communicate does not eliminate the prisoner's dilemma, since there may be various reasons why enforceable agreements cannot occur. In a fascinating example based on the opera Tosca, Anatol Rapoport (*Scientific American*, 1962) shows how the *fact* of an agreement does not change the basic prisoner's dilemma situation since it is in the interest of each party *not* to keep it.

The basic argument is that the actual outcome, an effort convention, would be suboptimal but a Pareto improvement over the *PD* solution. Because there are random elements in the determination of a convention, it is unlikely that the convention will coincide with the optimal effort solution $\pi_1 W_1$. Since conventions depend on history, it is likely that effort conventions that may have been close to optimal at an earlier period are no longer in that position later. Also, anybody who enters the firm will usually find certain effort conventions and sanctions designed to maintain them already existing. It is not myopia that deters individuals from moving from a nonoptimal effort convention to an optimal one, but the fact that it will rarely pay for any individual (in a fairly large firm) to try to organize all other individuals to put in more effort, and to negotiate all related changes.

Alam considers only the vertical relations —those between managers and other employees. Unfortunately, I did not give a sufficient treatment to the horizontal ones (intrapeer group) in my paper, although they are treated in my book (1976). In considering the choice options that involve simultaneous vertical and horizontal relations, we try to capture both group interests and individual interests. Because of effort discretion and *W* discretion, there are simultaneously 1) possible adversarial relations between different *groups* in the hierarchy, and 2) *individual* free-rider incentives towards low-effort levels, and lowering some components of *W*. Both the adversarial and free-rider forces lead to a *latent PD* problem, but the *PD* outcome is avoided by a (usually existing) effort convention solution.

Alam argues that I use a narrow definition of rational behavior. I believe this is not so. Rational behavior implies that each individual, or each unit, behaves in such a way so that they do as well as they possibly can. Where the payoff matrix can lead to a prisoner's dilemma, this implies that each side will do as well as possible for *any* choice made by the other side. If there is a strategy that is best irrespective of what the other side does, then this strategy is the rational strategy.[1]

---

*Professor of Economics and Population, Department of Economics, Harvard University, 204 Littauer Center, Cambridge, MA 02138.

An essential difference in Alam's and my views involves what can be achieved through the extraction of effort by threats and monitoring versus other methods. In my view, freely given effort results in greater output than through monitoring and threats. Thus, we can visualize effort as falling into two classes, including a lower level which involves the threat-monitoring approach, and an upper basically nonthreat level that is more effective. The assumption that I make is that the discretionary effort level is essentially the nonthreat, upper level—the firing point is at the point $A$ or to the right in Alam's figure. Hence, we view representative individuals using their discretion to choose effort levels up to a point, but not beyond, at which they would seriously risk being fired.[2]

For some analytical purposes, Alam's effort supply functions may be useful, but for the problem at hand there appear to be some difficulties: 1) there is no way for management to know the nature of the effort supply function; 2) if effort discretion exists, and I believe it will always exist to some degree, then it is unlikely that the firm can set $W$ and $p$ and expect to get $\pi_1$. Offering a higher wage rather than a lower one does not necessarily lead to more effort. Individuals would still be free to choose less.[3]

---

[1] In other game theory situations, maximization may not be meaningful since the individual agent does not control all of the variables, and there is no best strategy, irrespective of what the opposition does.

[2] These and related points are covered more fully in my recent paper (1983).

[3] If the employer controls both $W$ and $p$, then a higher $p$ is a substitute for a higher $W$. Employers would be motivated to *substitute* more $p$ for less $W$ whenever possible. But the attempt by employers to do so would surely be viewed as adversarial and, in my view, it would usually have effort-limiting consequences.

## REFERENCES

**Alam, M. Shahid,** "Intrafirm Productivity: Comment," *American Economic Review*, September 1983, *73*, 817–21.

**Leibenstein, Harvey,** *Beyond Economic Man*, Cambridge: Harvard University Press, 1976.

______, "The Prisoners' Dilemma in the Invisible Hand: An Analysis of Intrafirm Productivity," *American Economic Review Proceedings*, May 1982, *72*, 92–97.

______, "Property Rights and X-Efficiency: Comment," *American Economic Review*, September 1983, *73*, 831–42.

**Rapoport, Anatol,** "The Use and Misuse of Game Theory," *Scientific American*, December 1962, *207*, 108–18.

# The Neoclassical Maximization Hypothesis: Comment

*By* BRUCE J. CALDWELL*

In a recent paper in this *Review*, Lawrence Boland argues it is "futile" to criticize the neoclassical maximization hypothesis. His argument is summarized as follows.

1) "There are only two types of criticism of any behavioral hypothesis once one has established its logical validity. One can argue against the *possibility* of the hypothesized behavior or one can argue against the empirical truth of the premise of the hypothesis" (p. 1031). Boland hopes to show that neither of the two forms of criticism works against the maximization hypothesis.

2) One cannot *prove* the *logical impossibility* of the maximization hypothesis. Attempts to do so fail for one of two reasons: either one must assume that inductive proofs are necessary for true knowledge, or one must assume that true knowledge is necessary for maximization (see pp. 1032–33). Boland argues that neither assumption can be maintained.

3) Regarding empirical criticism, "The methodological problems of empirical refutations of economic theories are widely accepted" (p. 1033).

4) Even were this not true, the logical form of the maximization hypothesis (an "all-and-some statement") makes it neither verifiable nor refutable. "[E]ven if one could prove that a consumer is not maximizing utility...this would not constitute a refutation of the neoclassical hypothesis" (p. 1034), because one can always claim that the consumer was maximizing *some*thing, that is, something *other* than utility.

5) Finally, this does not imply that the hypothesis is necessarily a tautology. The hypothesis is best considered an untestable "metaphysical statement" which is an essential component of the neoclassical research program.

Points 3 and 5 above attributed to Boland are acceptable as stated. They are related points: the empirical problems mentioned in point 3 are sufficient to establish the untestability of the maximization hypothesis, which supports the conclusion in point 5. The purpose of this comment is to criticize points 1, 2, and 4. Specifically, it is shown that though there are only two forms of *logically compelling* criticism, many other forms of criticism that are useful exist; and that given a broader interpretation of the term criticism, Boland's remarks about proofs of logical impossibility, though correct, miss the point of what critics of the maximization hypothesis tried to do. An alternative to Boland's characterization of the maximization hypothesis as an all-and-some statement, one more consistent with well-known developments in philosophy of science, is also presented. In the final section, it is shown that Boland's approach of necessity neglects some widely known and relevant developments in the methodological literature.

## I. The Untestability of the Maximization Hypothesis

An area of agreement can be noted first: Boland is on firm ground when he states that the maximization hypothesis is untestable. (Metaphysical seems a rather loaded word, but it is acceptable.) There are a number of problems associated with testing the hypothesis; perhaps the most telling is that any direct test, including the revealed preference approach, requires that assumptions be made concerning the stability of preferences of the choosing agent, as well as the states of information confronting him. Since the content of these assumptions (properly called initial conditions) are subject to change but are not themselves directly testable, test results (whether confirming or disconfirming) are not unambiguously interpretable. Whether we interpret this to mean that the hypothesis is untestable, or that it is conceivably testable but that test results are not unambiguously interpretable, is a matter of semantic preference. The point is that the maximization hypothesis is neither directly (since preferences are nonobservable, and introspection and surveys are ruled out as unreliable) nor indirectly testable, in the usual senses of these terms.

---

*Associate Professor of Economics, University of North Carolina, Greensboro, NC 27412. I would like to thank Lawrence Boland and members of the New York University Colloquium in Austrian Economics, 1981–82, for many valuable comments on earlier drafts. Any remaining errors are my own.

## II. Forms of Criticism

Boland's assertion that there are only two forms of criticism of a logically valid behavioral hypothesis is true if one counts only logically compelling criticism as legitimate or important. But such a narrow definition of criticism is of little use when one considers the appraisal of scientific theories, all of which (if we take Kuhn and Lakatos seriously) contain untestable, "metaphysical" hard cores, theoretical terms which do not make reference to observable entities, and the like. In the evaluation and criticism of scientific theories, a number of criteria of appraisal may be employed, depending on the purposes of the theory in question: predictive adequacy, simplicity, generality, heuristic value, mathematical elegance, plausibility, and extensibility are among the criteria commonly mentioned. The definitions of, and the relative weights that should be attached to, such criteria have provided the grist for numerous debates in economic methodology. Criticism can also take place on another level —Karl Popper's distinction between internal and external criticism comes to mind.[1] In short, there are many routes to criticism in the appraisal of scientific theories; Boland's definition of criticism is overly narrow. His restricted definition causes Boland to misinterpret the writings of other economists, as is shown next.

## III. Proofs of Logical Impossibility

Boland is correct in asserting that no one has thus far *proven* the *logical impossibility* of the maximization hypothesis: after all, even if an agent has *zero* information, he still may choose the "best" option, or think that he has. But Boland's point is trivial: no one has ever attempted to prove the logical impossibility of the maximization hypothesis!

Boland is blinded by constructs which are of his own making. *If* one believes that there are only two forms of criticism, then one might believe that the works of Keynes, Hayek, and Shackle are attempts to prove logical impossibility. But a far more reasonable interpretation of these economists' efforts is that they all questioned the *usefulness* of a research program that either ignores the question of information, or posits agents with full information over all past, present, and future states of the world. Both Hayek and Shackle are subjectivists (as was Keynes on some readings), and *that* fact explains the emphasis on decision making under uncertainty, on the formation of expectations, and on the process of knowledge acquisition and dissemination that figures so prominently in their works. By focusing on proofs of logical impossibility, Boland misses both the intent, and worse, the significance of these economists' contributions.

In recent years, economists outside of the subjectivist camp (and this includes both advocates and critics of the neoclassical program) have asked similar questions. Indeed, even if one focuses only on the present "orthodoxy" in economics, such questioning has led to some of the most important recent work in economic theory-decision theory; various approaches to the modelling of expectations formation; the rationale for a monetary economy; search, contract, and optimal control theory; and so on. It is a telling indictment of Boland's definition of "criticism" that the sorts of questions that have led to these developments either would not count as criticism, or would be lumped with the straw man of "proofs of logical impossibility," and thereby dismissed as unfounded.

---

[1] Criteria of theory appraisal are discussed in Vincent Tarascio and myself. See Stanley Wong for an application of the internal-external criticism framework to a research program in economics.

## IV. The Logical Form of the Maximization Hypothesis

Boland notes that affirmative universal statements are not verifiable, even if true, and that affirmative existential statements are not refutable, even if false. He then argues that the maximization hypothesis, restated as "*All* consumers maximize *some*thing," is a hybrid all-and-some statement that is neither verifiable nor refutable, hence untestable. Though Boland's conclusion that the hypothesis is untestable is correct, it only works because the term "utility" has been replaced by the term "something" in the hypothesis. A more direct approach, offered below, demonstrates that the hypothesis "All consumers maximize utility" is untestable, *not* because of its logical form, but due to the presence of an undefined theoretical term, utility, in the sentence. This more direct approach has the added advantage of being consistent with well-known results in philosophy of science. The argument is by analogy.

Take the sentence, "All consumers maximize glitch." It has the logical form of an affirmative universal statement, so should be refutable, but not verifiable, according to Boland. But because "glitch" has been left undefined in the protocol language, the statement is neither verifiable nor refutable.

If the undefined term "utility" is substituted for "glitch," the same problem exists, *not* because the form of the sentence has changed, but because (in many formulations by economists) utility is undefined. Boland's logical objection, which holds when "utility" is taken to mean "something," can also be viewed as an empirical problem: a sentence containing an undefined theoretical term is not directly testable. (In their attempts to "operationalize" the concept of utility, revealed preference theorists avoid this problem. Unfortunately, they are still faced with the other empirical problems mentioned earlier.)

Boland's argument provides a textbook example of why logical empiricists since the mid-1930's have avoided discussing scientific theories in terms of the cognitive significance of individual sentences. Their predecessors, the logical positivists, tried to rid scientific discourse of the metaphysical by insisting that only analytic and synthetic sentences be accepted as cognitively significant. Identifying analytic sentences seemed to pose no problems (until Quine raised some questions), but it was necessary to find some criterion by which legitimate synthetic statements could be separated from metaphysical ones. Testability was the criterion chosen, but making that notion concrete proved difficult. Neither verifiability nor falsifiability were acceptable criteria because sentences expressed in certain common forms (affirmative universal and affirmative existential) would have to be ruled out as "metaphysical," even though they are widely used in scientific discourse.

An additional problem, independent of the notion of sentence forms, existed. Sentences making reference to unobservable theoretical entities (atoms, magnetic fields) often cannot be directly tested. Some philosophers and scientists (following Bridgman and Mach) recommended that all such theoretical terms be replaced by operational ones, but most recognized the important role played by theoretical terms in scientific theories.[2]

Both problems were resolved when strength of confirmation became the criterion of cognitive significance embraced by logical empiricists like Carnap, Hempel, and Nagel. Rather than check each sentence for cognitive significance, entire theories would be tested. Not every term would be required to have an empirical counterpart, but such undefined "theoretical terms" would gain meaning indirectly when the theories in which they were embedded survived numerous test instances. The logical empiricist solution has been challenged by both Popperian falsificationists and growth of knowledge theorists like Kuhn, but that is another story.[3] What is important for our purposes is that when utility is viewed as an undefined theoretical term, the hypothesis is rendered untestable, and that this approach avoids the circumlocution of transforming the hypothesis into an all-and-some statement.

---

[2] The necessity of retaining theoretical terms in scientific discourse is discussed in Carl Hempel.

[3] For a more thorough development of these themes, see my earlier article.

## V. The Ingenuity of Economists

It can finally be noted that many (though not all) economists have recognized that the maximization hypothesis is untestable, and that their responses to this apparent dilemma have been varied and, at times, ingenious.

Empiricists of all stripes try to reformulate the problem in operational terms. Certain Austrians, on the other hand, insist that the hypothesis (albeit expressed in slightly different form) is the fundamental axiom of human action which is known to be true a priori but which nonetheless has empirical content. (This is perhaps the most radical interpretation in that it presupposes an alternative epistemology.) Fritz Machlup views "homunculus oeconomicus" as an ideal type, an artificial device whose acceptability as a theoretical construct depends on the range of applicability of the theories in which he is used. James Buchanan distinguishes between a nonpredictive logic of choice, in which choices are based on the subjective perceptions of individual agents, and a predictive science of choice, in which specific, falsifiable predictions about choice behavior are made. Milton Friedman argues that the assumption should be ignored altogether, and that economists should focus instead on the predictive adequacy and simplicity of theories. Gary Becker and Armen Alchian insist that market behavior, not individual behavior, reveals rationality. And for those who believe that the individual agent should still receive attention, Becker joins with George Stigler to argue that alleged differences in tastes and preferences, usually viewed as exogenously determined by economists, can themselves be explained as responses to changes in incomes and relative prices.

Thus both advocates and opponents of the hypothesis have discussed at length the status of the maximization postulate. Sometimes the hypothesis is viewed as possessing empirical content, and sometimes it is expressed as a tautology; sometimes it is the centerpiece of a theoretical system, and sometimes a barely noticed artifact. Boland's contribution, though it sounds quite controversial, actually adds little to this well-developed literature. Only if one accepts his exceedingly narrow definition of the term criticism does it make sense to say that it is futile to criticize the hypothesis. Significantly, even in the absence of logically compelling criticism, debate over the merits of the hypothesis have yielded many important contributions, both in the methodological literature and in economics generally.

## REFERENCES

**Boland, Lawrence,** "On the Futility of Criticizing the Neoclassical Maximization Hypothesis," *American Economic Review*, December 1981, *71*, 1031–36.

**Caldwell, Bruce,** "Positivist Philosophy of Science and the Methodology of Economics," *Journal of Economic Issues*, March 1980, *14*, 53–76.

**Hempel, Carl,** "The Theoretician's Dilemma," in H. Feigl et al., eds., *Minnesota Studies in the Philosophy of Science*, Vol. 2, Minneapolis: University of Minnesota Press, 1958, 37–98.

**Tarascio, Vincent and Caldwell, Bruce,** "Theory Choice in Economics: Philosophy and Practice," *Journal of Economic Issues*, December 1979, *13*, 983–1006.

**Wong, Stanley,** *The Foundations of Paul Samuelson's Revealed Preference Theory: A Study by Method of Rational Reconstruction,* London: Routledge & Kegan Paul, 1978.

# The Neoclassical Maximization Hypothesis: Reply

*By* LAWRENCE A. BOLAND*

Bruce Caldwell has chided me for ignoring the contributions of modern philosophers in my 1981 article. Supposedly, if I were to appreciate the contributions of the "logical empiricists since the mid-1930's" I would see the error of my ways. Caldwell chooses to focus on my view of criticism—namely, that for any criticism to be successful it must be decisive, or as he says, "logically compelling." He feels this concept is "overly narrow."[1] I understand him to be saying that there are other concepts which are "less" narrow and thus my concept is only a special case. He recommends a more liberal concept of criticism which he calls "appraisal."

My reply is that he is exactly wrong. The idea that criticism must be logically compelling is not narrow but, instead, the broadest possible. And worse, his concept of criticism is completely inadequate. The implication that we should avoid logically compelling criticism in favor of a weaker line of argument (which he calls appraisal) is merely an expression of Caldwell's advocacy of the commonly promoted, but logically inadequate, methodology which philosophers today call "conventionalism."

## I. Appraisal as Criticism

From the standpoint of logic, it would seem to many of us that there is only one form of logically compelling criticism—namely, the demonstration that the argument being criticized leads to a contradiction and thus is logically invalid. Note that this was not the basis of my argument in my earlier paper. My paper was about arguments about the neoclassical maximization hypothesis that arise *after* "one has established its logical validity." That is, the maximization hypothesis, like any hypothesis, asserts that if certain prior conditions are met, then, *necessarily* particular subsequent conditions will be met. In short, if the prior conditions are all true then the subsequent conditions will also be true (see further, my 1979 article). Logical validity concerns the term "necessarily," and thus if logical validity has been established, the hypothesis can only fail to explain the truth of the subsequent condition because one or more of the prior conditions cannot or are not met. The former is about the possibility of the hypothesis being employed in a successful explanation and the latter is about the empirical truth of an explanation based on the hypothesis.

I will have to agree that even these two forms of criticism can be seen to be invocations of logically compelling criticism. Any condition is truly impossible only if it leads to a necessary contradiction (a classic example is Kenneth Arrow's Possibility Theorem), and a condition is empirically false whenever the conjunction of it with an observation statement necessarily yields a contradiction. So, if Caldwell is correct, there must be some other form of criticism which can never be reduced to a claim that there exists a contradiction.

What alternative does Caldwell offer? First, he gives us a list of conventionalist criteria—simplicity, generality, mathematical elegance, plausibility, and so forth. Presumably, such criteria are to be used to "appraise" economic explanations in the same way a welfare function is used to appraise alternative economic policy recommendations.[2] But how are such criteria used to form a criticism? Perhaps one is supposed to adopt a criterion and use it to measure a given model or theory. Now the only possibility of a criticism that can be advanced against the given theory is that *if* one's aim is to maximize according to the accepted criterion, then the criticized theory somehow fails to achieve the maximum. But failure to achieve one's aim can easily be reduced to a failure to avoid a contradiction between one's aim and one's achievement. In other words, Caldwell's first alternative form of criticism, the use of conventionalist criteria, does not avoid the use of logically compelling criticism.

---

*Department of Economics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6.

[1] This kind of discussion can get quite awkward. Is his criticism of my paper intended to be logically compelling, or something else? For my purposes I will assume that he is intending to convince us of something —in this case, it is that my concept of criticism is too narrow.

[2] In effect, the methodology of appraisal is a variation on neoclassical analysis and it is for this reason I examined methodology in these terms (see my 1971 article). The difficulty is that the use of such criteria in the appraisal of theories or models fails to fulfill its objective for the very same reason that plagues welfare economics—there does not exist a universal criterion

His second alternative, "Popper's distinction between internal and external criticism," fares the same fate. Internal criticism is merely based on the theorist's acceptance of his or her aims and failure to achieve them. External criticism centers on criticism of the theorist's aim by measuring it against some externally given criterion. In either case, my above discussion of aims applies, but with a more general view of what constitutes an aim. Namely, his second alternative involves criticism that a given explanation fails to solve some particular theoretical problem (Stanley Wong, 1978, pp. 23–24), and thus does not avoid being reducible to logically compelling criticism.

All forms of criticism depend on aims, criteria, testing conventions, and so forth, which are put beyond question *for the purposes of the criticism.* Any criticism succeeds only by showing that to remain consistent, the only alternative is to give up one's aims, criteria, and the like, to avoid being inconsistent. So, contrary to Caldwell's arguments, my concept of criticism is not narrow, but rather it is the most general, since all forms of criticism can be reduced to matters of logical consistency.

## II. The Poverty of Conventionalist Methodology in Economics

The only contribution of "logical empiricists [of] the mid-1930's" was to deflect interest from the difficult question of the empirical truth or falsity of explanations to the more convenient question of the logical validity of argument formed by any explanation. Rather than arguing about whether a theory is true or false, we are supposed to choose between the available theories by some criterion such as "simplicity" or "mathematical elegance." Instead of claiming that a theory is true or false, we are supposed to judge it as being better or worse than any other theory according to the accepted criterion. Obviously, this approach to methodology leads to an infinite regress (see further, my 1974 paper). By what supercriterion do we choose the best criterion?

Each of the criteria listed by Caldwell has its advocates and critics. No logical empiricist would ever claim that his or her employment of a supercriterion to choose a particular theory constitutes a proof that the theory is actually true. But if it is not intended to be a proof, what is accomplished? Or better still, how do we know when any critical appraisal is successful and when it is not?

The idea that we should appraise rather than criticize the theories of economists is promoted because most people feel that decisive criticism fosters unproductive fights and controversies over the truth or falsity of theories. Supposedly, reasonable people would see that theories cannot be proven absolutely true or false. And thus, to be productive we should be more tolerant. The difficulty with this "reasonable" view is that it fails on its own terms. The statement that theories are not true or false is itself a statement—one which is claimed to be true!

Some economists may feel better by advocating appraisal rather than logically compelling criticism, but it only postpones the arguments as we would still have to decide which appraisal criteria are true in order to make our appraisals. The room for differences of opinion and controversy is even greater when it comes to choosing one's criteria.

## REFERENCES

**Boland, Lawrence A.**, "Methodology as an Exercise in Economic Analysis," *Philosopy of Science*, March 1971, *38*, 105–17.

______, "Lexicographic Orderings, Multiple

---

that will work in all cases. A more general criticism of conventionalist theory-choice criteria—namely, that it is self-contradictory—is offered in my book (1982).

# Property Rights and X-Efficiency: Comment

*By* HARVEY LEIBENSTEIN*

Louis De Alessi's paper (1983) begins in an ecumenical spirit and suggests that property rights theory and the X-efficiency approach have not drawn on each other's contributions. I share that spirit. There is much to learn from a variety of approaches. But, as De Alessi proceeds, the ecumenical spirit withers. The last three sections argue to the effect that general neoclassical theory (i.e., property rights theory)[1] obtains the same results as X-efficiency theory, but has a superior methodology.

De Alessi's paper puts me in the position of a reluctant dualist. His approach resembles an intellectual takeover bid since it argues 1) that neoclassical theory is flawed, 2) that the property rights approach has corrected the flaws, and 3) by implication that it supercedes other micro approaches.[2] The concepts of property rights, transaction costs, and adjustment costs, and their motivational implications, are certainly useful. I see no reason to argue against them, or against property rights theory. But I believe that there is more to the explanation of inefficiency. Thus I will stress the special nature of the X-efficiency approach, examine our differences, and attempt to correct some of De Alessi's misinterpretations. Nevertheless, I believe that De Alessi's article renders an important service since it gives a succinct summary of property rights theory and explains its significance.

The main differences between us lie in the basic presupposition behind the two approaches. In answering the motivation question, property rights theory seems to say that property rights are the only motivating force, and that people act to fully maximize wealth given the limits of their property rights. When they do so, their activities are efficient. The X-efficiency approach raises a different type of question. Namely, it asks (*within* the firm and *outside* of the exchange relationship) what are the motivating forces that determine effort and productivity, including the motivating forces leading to maximizing or less-than-maximizing behavior. Property rights are, of course, a motivating force. However, it is one of many, not the only motivating force, nor in many instances is it the most important one determining effort. It is important to highlight these differences because different modes of analysis determine not only results, but what the analyst looks for and possibly finds. The value of a theory depends in part on the research it generates.

Another difference lies in the interest in *inefficiency*. It is not clear whether De Alessi cares about inefficiency. The postulate of maximization of utility by all individuals and the explicit assertion by De Alessi that "The equilibrium solution associated with a given set of constraints is efficient" (p. 73) gives the impression that inefficiency is assumed away. Clearly, the X-efficiency approach sees inefficiency as a major problem. Only if one employs an approach that at least recognizes within it the possibility of inefficiency can one have empirical results which indicate an approximation to efficiency. Inefficiency must be a variable if we are to be able to find meaningful empirical instances which are efficient.

Compared to the neoclassical approach, the property rights approach seems to dissect the firm into component individuals and their implicit contracts. After dissection, the property rights approach assumes that the ordinary utility-maximization rules still apply.

---

*Professor of Economics and Population, Department of Economics, Harvard University, 204 Littauer Center, Cambridge, MA 02138. I am indebted to H. Barkai, J. Dean, R. Frantz, E. Helpman, J. Medoff, and J. Rothenberg for valuable comments. They are not responsible for errors or viewpoints.

[1] The labels "property rights theory" and "general neoclassical theory" will be used interchangeably.

[2] To stake his claim, De Alessi might have considered at least some other alternative approaches, such as the work of Oliver Williamson (1970) and that of Richard Nelson and Sidney Winters (1982).

For the X-efficiency approach, this is not the case. Once *inside* the firm, *outside* of markets, it is strategic decision making that becomes critical; individual rationality no longer implies collective rationality.[3] I hope to show that nonmaximization inside the firm becomes an important element in motivating effort and in solving free-rider effort problems.

## I. Purposes, Separability, and Behavioral Focus

While a theory may have a variety of purposes (such as an aid in the description of behavior, or in the analysis of some set of circumstances,[4] as well as making conditional predictions), it may help if we view X-efficiency theory as a vehicle for the analysis of some types of nonoptimal behavior. Its domain focuses on those areas which do *not* involve the phenomenon of exchange. Thus, we might visualize three types of inefficiency: 1) inefficiencies handled by standard theory which involve faults in the market; 2) a second type involves nonoptimal economic behavior *within* established units such as the household and the firm, *apart from the buying and selling of goods*; and 3) a third type concerned with the interaction of the market and intraunit behavior.

The X-efficiency theory deals with units within which exchange *does not* take place, while neoclassical theory and general neoclassical theory (i.e., property rights theory) do not allow us to consider the behavior within such units *in an essential way*. If we are concerned with the possibility of inefficiency apart from exchange, then, of course, we must use a classification that will allow us to analyze behavior entirely apart from exchange contracts.

Consider the following thought experiment. Suppose a firm has two divisions: a contract ($C$) division for arranging purchase and sales contracts and a production ($P$) division. Suppose that $C$ and $P$ are in separate locations. All buying and selling contracts are arranged by $C$; all other activities by $P$. Surely we can speak of the different efficiencies of the $C$ and $P$ divisions. Inputs already exist for $P$. Given the inputs, $P$ produces the outputs. It then informs $C$ of the outputs during the week and $C$, say the following week, sells the outputs, and orders more inputs to be delivered to $P$. Surely there is a distinction to be made between how well $C$ carries out buying and selling, and how well the production division carries out the input transformation activities. Thus it makes sense to speak of the effectiveness or efficiency with which inputs are *used* in order to create outputs. In this example, there is complete separation between the contracting for inputs and the using of the inputs to create outputs. But, the essence of the problem does not change where the separation is not quite so clear cut.

In standard theory, the X-efficiency problem is obscured (actually eliminated) by the introduction of a production function which connects inputs to *maximum* outputs. Given prices, the production function enables the calculation of profitable exchanges where one side of the exchange involves the purchase of inputs and the other the sale of outputs. But if we want to study the possibility (and the actuality) of nonoptimal behavior in the production department, then we cannot employ a production function whose dependent variable is maximum output.

It seems reasonable to interpret property rights theory as one which uses a complex production function.[5] In other words, it is a theory that allows for monitoring activities, transaction activities, and adjustment activities, and the cost associated with all these. Some employees (managers) monitor the activities of other employees. Similarly, some of the employees' activities involve the cost of transactions; others somehow involve the cost of incompletely allocated property rights. Since all the activities are carried out in an optimal fashion, the whole thing really amounts to a somewhat complex production function that translates inputs into optimal outputs, *given the constraints*. Clearly, such a "production function" does not allow us to *see* behavior which allows for the possibility of nonoptimal results.

---

[3] On the distinction between the two, see Anatol Rapoport (1974).

[4] To my mind it is an oversimplification to view theory as having only a single purpose, for example, prediction.

[5] See, for example, Michael Jensen and William Meckling (1979, p. 471).

In X-efficiency theory, the production division exists in an essential way. Indeed, it is the locus within which activities take place to transform inputs into outputs *at various possible rates*, many of which are nonoptimal, since both firm members classified as managers as well as other employees have discretionary spaces within which to choose activities, that is, they work with incomplete contracts. Such discretionary spaces may exist with respect to people's effort, the use of capital, the use of information flow, access to information, and so forth. The boundaries of discretionary spaces are not always clearly recognized. Such differences can readily generate adversarial feelings between individuals at different levels. Furthermore, conflicts over discretionary space are not necessarily zero-sum games; they may be negative-sum games. The production division determines what the outcome is within a variety of discretionary spaces.[6] In essence, two different classifications are involved between X-efficiency theory and property rights theory. De Alessi always connects the (contracted) utilization of inputs with exchange; in addition, it is assumed that people always behave optimally. Therefore, it is not surprising that our approaches and conclusions differ.

## II. Decision Making

De Alessi argues that I assume nonmaximization, but that, somehow to De Alessi, it amounts to maximization. What I really postulate is a relationship between pressure and "calculatedness" (with "full" calculatedness approximating maximization), so that both nonmaximization and maximization are included as possibilities. By pressure, I have in mind the sum of particular pressures on an individual, including those that arise (a) from within the individual, (b) from within the firm and applied by others on the individual, and (c) those from outside the firm (for example, competition from other firms) and transmitted through the firm on intrafirm decision makers. Now, we can visualize pressure on the $X$ axis and the quality of decision (degree of calculatedness) on the $Y$ axis. As pressure *increases*, the quality of decision making increases until some optimal point (or set of optimal points) is reached. The optimal points are equivalent to maximizing behavior.[7] Beyond the optimal points, we may have too much pressure and suboptimal decision making sets in. This relationship is similar to, or the same as, the psychological law known as the Yerkes-Dodson law which has been validated many times in a variety of contexts since its enunciation in 1908. To many the law seems "obvious," since it accords well with experience.

To clarify what I mean by nonmaximization, it helps to indicate types of possibly nonmaximizing behavior. Let us look at decision making on the basis of Herbert Simon's (1978) distinction between a substantive and a procedural theory. A substantive theory states the objective (for example, maximizing utility) without indicating how it is done, while the procedural one is concerned with specific procedures or some general statement about the procedures. Now, there is no established list of procedures for maximizing utility. Some of the following come to mind: 1) determination of the entities in the objective function; 2) the internal ordering of tastes; 3) obtaining external information of "goods" and other options available; 4) external information of prices or opportunity costs; 5) determining the constraints faced by the individual; 6) determining when to stop information search; and so on. The point of all this is not to suggest a complete or practical list,[8] but to suggest that there is a lengthy group of procedures, and that there are a variety of possibilities of not completing them. Hence, we can think of *incomplete sets of procedures*, or of sloppy or imperfect ways of following some procedures, as forms of nonmaximization.

---

[6] See Robert Smiley (1976) and Williamson (p. 100) for estimates of the discretionary spaces the stock market allows firms.

[7] I regret that one of my articles (1982) referred to here and in the next section were not available when De Alessi wrote his paper. I feel I have to restate, however briefly, some of these ideas in order to clarify the differences between De Alessi and myself. On the Yerkes-Dodson law, see J. W. Atkinson and O. Birch (1978) for a general treatment. I am indebted to David C. McClelland for pointing out the close connection between some of my ideas on decision-making behavior and the Yerkes-Dodson law. All of this is not to deny that some maximization models may work well for some purposes.

In addition to 1) incomplete procedures or faultily executed procedures, I can think of 2) habits, 3) conventions, 4) Simon's satisficing, 5) the commitment to an ethic or objective which leads to losses that could otherwise be avoided, and 6) forms of sacrificing behavior, as types of nonmaximizing behavior. The point to be emphasized is that the pressure-decision relation is a generalization that includes both maximizing and nonmaximizing modes of behavior. I will show below that some types of nonmaximization behavior are useful in helping us obtain reasonable solutions to some strategic decision-making problems.

It is of special interest to observe that to interpret the meaning of maximization as inclusive of the disutility of making decisions leads to contradictions. Consider a utility function $U(G, D)$ containing two arguments; $G$ for goods and $D$ for decision procedure. Under $D$ there are two options, a maximizing and a nonmaximizing procedure. Now suppose that the decision maker takes a smaller $G$ because he does not like to carry out the decision procedures he would have to use to get a larger one (i.e., maximization). In other words, his total utility is composed of the utility of the good plus the utility of nonmaximizing behavior. *Thus this individual would be said to be maximizing when engaging in nonmaximizing behavior.* This is an obvious and well-known paradox of a class of self-referential situations in logic. Obviously (or perhaps not so obviously), it makes more sense to start out with the recognition that nonmaximizing decision behavior exists and to refer to it as such. Paradox lost is clarity gained.

"On-the-job leisure" is a troublesome side issue that results in some misunderstandings. The essential argument (which De Alessi restates) is to the effect that X-inefficiency does not exist, but involves a misspecification of the product. If effort is lower rather than higher, is more on-the-job leisure produced? The basic question is whether an individual receives less satisfaction from his work qua work when he puts forth more effort. In my view, this depends on the nature of the work, whether others are working equally hard, and whether approval is associated with the marginal effort. Starting with a low effort level and continuing towards higher levels, consider four phases: 1) a positive marginal utility phase because of boredom at low effort levels; 2) a hard-to-calibrate portion where approximately constant positive utility prevails; 3) a declining but *positive* marginal utility phase; and 4) beyond that phase a negative marginal utility for effort so that more effort is associated with less satisfaction. Thus there may be circumstances (the first three phases) where an increase in effort is not associated with any loss in satisfaction. In other cases there is such a loss, and other things equal, the incentives for individuals are towards putting forth less effort.

In general we should keep in mind that effort satisfaction is likely to depend on whether relevant others are putting forth more or less effort, as well as the simultaneous approval or disapproval of others associated with marginal effort. Thus there may be a phase where from an individualistic viewpoint more effort involves less marginal utility, but when the effort behavior and possible approval of others is taken into account, then these elements may counterbalance what would have been a loss of marginal utility if the individual had acted in isolation.

## III. Game Theory and Effort Conventions

Since the possibility of suboptimal outcomes is one of the issues, I consider briefly, in the next paragraphs (without any of the detail), how such outcomes may arise. Ideas previously published will be considered very briefly, but some aspects treated lightly elsewhere, and especially relevant to some of the issues, will be considered more fully. The essential idea is that intrafirm interactions involve both cooperative and competitive (or adversarial) aspects. The existence of discretionary spaces controlled by a number of decision makers simultaneously implies that no one controls all the variables. Hence the choice problem is a game theoretic one. The basic argument is that there are (a) conflicting interests between representative individuals in different hierarchically placed groups; and (b) simultaneous free-rider effort options for peer group members. Both incentives tend toward a *latent* prisoner's dilemma (*PD*) solution, but (c) the *PD* outcome is avoided by the *existence* of (usually) suboptimal effort conventions.

---

[8] Such lists do exist. Many are much more extensive than the above. See especially Alexander Cornell (1980).

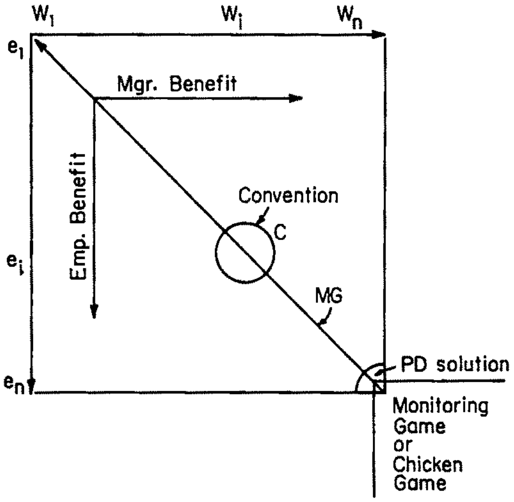


FIGURE 1

With the aid of Figure 1, consider adversarial behavior between two levels in a hierarchy, say employees and managers. Employees have some degree of effort ($e$) discretion while managers have discretion with respect to wages and working conditions ($W$). Employees' effort options are ordered as follows: $e_1 > e_2 >,\ldots, > e_n$; that is, $e_1$ is the greatest effort level. Managers have wage and working condition options whose costs and value to employees are as follows: $W_1 > W_2 >,\ldots, > W_n$. The less $W$ for any effort level, the higher the profits. Now for any given effort level, the firm (managers) would want to move towards $W_n$ (which involves the lowest cost per man). The representative employee in his relations to superiors usually has the incentive to put forth more effort towards his personal interest (up to a point, say the point $e_n$) and less in the firm's interest. For any given level of $W$, employees want to move towards $e_n$. Thus, the incentives in the system lead to a latent prisoner's dilemma result if both sides attempt to operate as "maximizers," that is, to do as well as they can irrespective of what the other side does. Simultaneously, we have to keep in mind that *within* each hierarchical group, there are free-rider incentives; that is, every employee may wish to work as little as possible even though he may desire that others work reasonably effectively so that the enterprise and his job continue. These free-rider incentives exist at every level. Even if some exceptionally wise management eliminated the adversarial feelings on the part of employees, their free-rider options would still provide incentives towards a latent *PD* outcome.

However, the *latent PD* outcome does not normally occur. Usually, conventions exist which determine the effort level and solve not only the adversarial problem between two *groups*, but simultaneously solve the free-rider incentive problem for individual members of each group. Managers accept the effort convention and pay accordingly, that is, according to a wage and working condition convention or according to pay for somewhat similar work elsewhere. In Figure 1, the effort convention is shown by the letter *C*. The circle suggests the inert area bounds around the convention. As illustrated, the effort convention is superior to the latent *PD* outcome but not as good as the optimal outcome $e_1W_1$. Points along the line *MG* show Pareto improvements as we move in the direction of the arrow. Since higher points on *MG* involve more effort, more productivity, and higher wages, it is possible to have a distribution that would benefit both employees and yield higher profits. There are many other "diagonals" roughly parallel to *MG* which would also involve Pareto-improvement loci. Note also that once the convention exists, only the effort convention is observable and there is no necessary awareness of the latent *PD* problem.[9]

There are two aspects to conventions: (a) a convention is a regularity of behavior out of a set which includes two or more possibilities, but only one such regularity can exist at any one time; and (b) sanctions supporting the convention are frequently applied when deviant behavior goes beyond certain bounds. It is assumed that the *effort* convention is based on history, sanctions, and the observation of the average effort norm set by others. Thus each person will choose an effort level depending on what he learns about the way things have been done or are being done, signs of approval or disapproval, especially from peers, and his estimation of the observed effort norm. An equilibrium effort level is established if, on the average, the supply of effort of the sum of individuals is equal to the observation of what the average level appears to be.[10]

There are a number of elements to be noted about effort conventions. The convention operates more or less on the basis of a stimulus-response mechanism once it exists. Given the context, people usually react by employing the conventional mode of behavior.[11] Are effort conventions stable? Would people stick to a nonoptimal convention or attempt to move it to a Pareto optimal one? The costs, where a fair number of individuals are involved, are likely to be a significant deterrent for either maximizers or nonmaximizers. Among the cost elements are the gathering and distributing of information, persuading others of the desirability of change, the cost of facing sanctions by strong adherents, the cost of persuading peers *and* adversaries, and so forth. Given these disincentives, it is likely that very few individuals will try to create a shift from a nonoptimal to an optimal convention.[12]

For some purposes, it may be useful to distinguish neutral, lower-bound, and upper-bound sanctions. Where the sanctions are neutral, they are applied to keep effort roughly between some upper and lower bounds. Those who have a lower-bound focus apply sanctions only on deviant behavior below the lower bound, while followers of an upper-bound focus apply sanctions on those who put forth too much effort. In the former case, people are concerned with others not "pulling their weight," while in the latter the deviants are viewed as "rate busters." Thus some systems of sanctions (with an upper bound focus) may permit the gradual erosion of the effort norm as new people enter the firm and perform at a somewhat lower effort standard than those that leave.

Relaxing the maximizing assumption makes it easier to accept and believe that the effort convention will override the *PD* solution. The stimulus-response mechanism makes it easier for each individual to believe that others will choose the same way. Furthermore, a nonmaximization choice procedure such as the emulation of peers, or a commitment to peer group behavior patterns, makes it easier to believe that few people will flout the convention even if there is a calculated advantage in not sticking to it. Since the convention involves imposing sanctions (which may be unpleasant), it is easier to expect that some will impose sanctions if such behavior arises out of emotional forces rather than cooly calculated considerations.

---

[9] The game theory approach to effort supply is considered in my 1976 book, ch. 9. The importance of conventions, though that term is not used, is considered in ch. 5. A well-developed formalized theory of conventions exists. See especially the book by E. Ullmann-Margalit (1977) for the most explicit treatment of conventions as solutions to *PD* problems. See also the work by D. Lewis (1969) and Andrew Schotter (1981). On the repeated *PD* game problem, see Robert Aumann (1981).

[10] In a diagram, an equilibrium would be established where the effort supply curve would cross a 45° line and where average effort supplied is a function of observed effort and history.

[11] Some conventions are intrafirm routines, but others may not be. The stimulus for the use of the convention may occur in a nonroutinized way. For an extremely interesting discussion of the role of routines in firms, see Nelson and Winter, pp. 96–138.

[12] The effort convention can be destabilized if some intrafirm or environmental variables (i.e., payoffs) go beyond the inert area bounds. Such shocks may lead to a new effort convention; but because of past history, past or existing adversarial relations, or a general lack of trust, the new convention need not, and probably will not, be optimal. It would take us too far afield to even speculate on the process of new convention creation.

Property rights represents a recognized way of doing things. It is a regularity of behavior and hence a convention. Like other conventions, it is strengthened by some types of nonmaximizing behavior. The costs of protecting property rights, especially private property rights, would be enormous if we could not count on the fact that others respect such rights, even when these others could gain by not respecting them. The essence is mistrust vs. trust. In a world where no one trusts anyone else, everyone expects their goods to be taken unless privately protected. Everyone would have to try to monitor everybody about everything constantly. If you hire monitors, who would monitor the monitors? Kenneth Arrow echoes these concerns when he argues that, "Collective undertakings of any kind, not merely governmental, become difficult or impossible not only because *A* may betray *B* but because even if *A* wants to trust *B* he knows that *B* is unlikely to trust him. And it is clear that this lack of social consciousness is in fact a distinct economic loss in a very concrete sense..." (1974, p. 26). Douglass North makes an important related point to the effect that, "Equally fundamental is the dilemma that people do not always 'free ride,' ...That is, if people are convinced the rules are fair, they may obey them even when at times they could be better off not obeying them..." (pp. 163–64).

It is worth noting that an effort convention within the production setup may be viewed as playing the same role as price does in competitive markets. The price becomes a *parameter* under competition which solves a coordination problem. Similarly, the convention serves as a parameter with respect to the coordination of effort since it permits different individuals to "rally around" the convention and to use it as a basis for their individual effort decisions. In other words, conventions solve game theoretic problems by "parameterizing" the environment for individual decision makers.

## IV. Monitoring and Effort

The impression one gets from property rights theory is that the degree of effort depends on an adequate amount of monitoring. The view taken here is different. While monitoring against possible shirking is one way of handling effort, it is unlikely to be an optimal way. There are two reasons for this: (a) monitoring implies a preset effort standard and hence a fairly complete effort contract; and (b) beyond some point monitoring itself is likely to have undesirable motivational effects on effort, especially on the qualitative components of effort.

Most employees are likely to have more information about the details of their job than the manager. Thus management's attempt to work out a fairly complete contract for labor in order to set an effort standard is likely to be a contract for a suboptimal effort level. For instance, it is well known that in many industries the best way to reduce output is to "work according to rule." Also, and equally important, is that an arrangement that emphasizes the monitoring-antishirking aspects of supervision is likely, by its very nature, to set up or increase adversarial feelings and relations between superiors and subordinates. Such relations in their turn are likely to reduce effort. Thus the question is not whether managers are residual claimants and as such sufficiently motivated to do adequate monitoring, but whether the interaction between managers and other employees (as well as between peers within the two groups) generates optimal attitudes. It is to be noted that at this stage the argument is not whether the view taken by X-efficiency theory is correct or not, but rather that it is different than the argument put forth by De Alessi in which the outcome depends on property rights motivating adequate monitoring.

What fairly extensive monitoring essentially does is to set up something that approximates a mutual threat game. The firm stresses that if a minimal (or agreed upon) level of performance is not achieved, then the individual will be punished; for example, fired. The individual employee, and employees as a group, on the other hand, are in a position to impose some damage to the firm by lowering effort, or strikes, and/or by contributing to labor turnover costs. The ideas are illustrated with the aid of Figure 1.

View $e_n$ as the highest fully *monitored* effort level. Below $e_n$ the employer suffers a productivity loss and the employee gets fired. Thus a cooperative solution of $e_nW_n$ is the best solution for all parties concerned under a monitoring-antishirking system.

On the other hand, the situation may be such that each side may feel that it can gain by a more extreme strategy if the other side does not resort to its extreme strategy; for example, a slowdown or a strike (for some improvement) versus layoffs or being fired. The unfortunate outcome for both sides is if each uses his extreme strategy. This approximates a "chicken" game. Most of the time both sides will try to avoid the extreme threats and choose $e_nW_n$. In Figure 1, the possibilities under a monitoring-threat system are shown in the area marked "monitoring or chicken game."

Of course, real labor relations "games" are much more complex. Not all managerial systems can be captured by effort conventions. In some cases a monitoring-threat system may be a superior alternative to a low-level convention (say one under which the effort level was below $e_n$), but, of course, that is a nonoptimal level. Another possibility is that monitoring may be combined in some way with an effort convention that exists so that what is essentially monitored is the convention itself.

The main general point is that merely obtaining an acquiescent nonshirking effort is of limited value. Freely offered effort, inclusive of attentiveness and caring about the quality of effort, in return for what is viewed as a good deal (in the long run) is likely to result in higher productivity. There exists a world of other options involving higher effort levels than $e_n$ for which the firm does not employ monitoring to any extensive degree, and instead resorts to other motivating forces, which in essence involve higher levels of trust and lower implicit adversarial relations. But such relations are likely to involve quite a bit of discretion on all sides.[13]

The basic argument is that for the same allocation of property rights in firms *A* and *B* (assuming a bounded definition of these rights), if *A* and *B* have different histories and different effort conventions have developed, then with exactly the same knowledge, two very different outcomes will result. A case pertinent to the above was reported in the *New York Times* (Steven Rattner, 1981), which discussed two Ford plants (in Germany and the U.K.) identically designed, and equipped to produce the same quantity of an identical product. The German plant produced 50 percent more automobiles with 22 percent *less* labor. Rattner cites Ford officials in the U.K. to the effect that, "The difference between the two plants comes down to the attitudes of the workers," and the operations manager of the German plant as, "It has to do with the mental attitude of our people here" (p. D1). Clearly, in this case, property rights is not the only motivating force.[14]

## V. Major Differences and Minor Issues

Sections IV and V of De Alessi's paper consider a large number of issues, some of which are important and some seem to be of a minor nature. I cannot respond to each and every issue. Instead, I will focus on a few of the main issues on which we seem to differ.

---

[13] Strictly speaking, I cannot prove that the view taken here is correct. However, my impression is that the organizational behavior literature supports this view. In general, the organizational behavior literature suggests that allowing for discretion and the participation by employees in decisions leads to more effective results than detailed direction or rules and close supervision. See Robert Miles (1980, pp. 93–99; 106–14). A personal note may be appropriate here. Recently I spent six months in Japan and visited some large firms. In all cases I inquired about the way in which employees were monitored. The managers were extremely sensitive to the idea and convinced me (and my Japanese colleagues) that no monitoring by superiors was attempted. Promotions, by and large, were a consequence of age and duration of employment. Here we have examples of some highly X-efficient firms who did everything possible to avoid monitoring. The managers were aware that such practices were carried out in the United States, but these were practices that they assiduously chose *not to use*, despite the fact that at some earlier period they were tried by some Japanese firms and dropped.

[14] Some studies by C. F. Pratten (1976) on productivity differences in plants in different countries belonging to the same company are pertinent. Some of the cases do not apply because of different size runs or different product mixes, but others offer rough approximations to the Ford Story. See especially, Appendix 4, pp. 83–105, Case Nos. 4, 5, 10, 12, 27, 33, 40, 62, 69; Appendix 5, pp. 106–114, Case Nos. 6, 21, 26, 34, 41.

A surprising argument in De Alessi's paper is the assertion that I believe that ownership form does not matter. My acceptance of the result of the Shelton "experiment," as well as the acceptance of other evidence of this type in my work, indicates quite the reverse. This point is very clear in the chapter on state-owned enterprises (ch. 8) in my 1978 book.

In De Alessi's Section IV, he argues that the postulates of X-efficiency theory (where they differ from property rights theory) can be obtained as results of property rights theory, and since property rights theory has fewer axioms, it is superior. A few examples will show why X-efficiency postulates cannot be obtained from property rights theory. Let me restate very briefly two of the postulates in question. 1) *Nonmaximization*: This is part of a more general decision postulate which states that nonmaximizing behavior will take place when pressure is relatively low or high. Selective rationality (see my 1976 book) allows for both maximization and nonmaximization. 2) *Inert area*: By an inert area, I have in mind a "position" (value of a variable) of an economic agent which does not change even if certain functionallly related independent variables do change, but stay within certain bounds.

De Alessi argues that a utility-maximizing individual does have the incentive to respond selectively, which means possibly nonmaximizing behavior. If transaction costs are sufficiently high, then an individual will respond only to those events that are sufficiently large to justify a change. This goes counter to any reasonable interpretation of what I have attempted to say. Nonmaximization does not depend on sufficient transaction costs; it is possible with zero transaction costs. For example, some individuals might not inquire about price even though this would be a costless inquiry, and hence not be in a position to make a maximizing decision.[15] Even if there are transaction costs, it does not mean that nonmaximizing behavior is implied. There is no necessary connection. Transaction cost may exist; the benefits of a move from $A$ to $C$ may be a maximizing move, while a move to $B$ may result in a gain but not a maximum gain. However, an individual moving according to habit, convention, or an incomplete decision procedure, may move to $B$.

A similar argument holds with respect to inert areas. It is true that if there are enough adjustment costs or other costs, people would not move. That is not the meaning of an inert area. It is postulated that people do not move from their present positions within the inert area bounds apart from such costs. It is true that you can always stick in enough transaction costs, adjustment costs, or attenuation of property rights to get the results you want once you know the results desired.[16]

Again De Alessi argues that contracts are incomplete because of monitoring costs. It is not at all clear that monitoring cost is the relevant consideration. Monitoring cost, and the anticipated gain from monitoring (if a gain is anticipated), need determine only the amount of monitoring to be used. In principle we could have complete contracts or incomplete contracts and zero-monitoring costs. For our purposes, it is quite sufficient to start with incomplete effort contracts as a basic postulate.

---

[15] I frequently ask students if they ask their dentist the price he will charge for what he is about to do. A fairly high proportion indicate (a) that they do not, (b) that it would be easy to inquire, and (c) that the bills every so often involve surprises. I believe it is reasonable to view such behavior as nonmaximizing. Some may be tempted to invent "possible reasons" for such behavior so that it fits maximization, even though they do not know the people or the facts. I assume throughout that we use "maximizing utility" in a nontautological sense.

[16] On inert areas, De Alessi seems to approve of my exposition. But here I have changed my mind. In recent papers I have simply assumed the existence of inert areas. In the past I have tried to argue that the cost of moving from one position to another was an *intrapersonality* "cost," which can lead to suboptimal behavior, which differs from an externally determined cost of moving, and which can exist apart from external cost. But this was misunderstood. It seems to simplify matters just to assert inert areas and inertial behavior within bounds as a postulate and not speculate on the reasons why.

A methodological issue between us, according to De Alessi, concerns *ex ante* predictions versus *ex post* confrontation of theory to empirical results. But the *ex ante, ex post* distinction does not really get at the roots of the issue. If one is to test a theory by making a "prediction," then the result appears to be more convincing if the prediction is a consequence of an inference from the theory, and the result is not known when the prediction is made. This has to do with effect rather than scientific significance.[17] There is room for much *ex post* analysis of data. Economic historians could hardly perform their tasks if they could not apply their theories to existing data. A major purpose of theory is to *account* for the facts, obtained either *ex ante* or *ex post*.[18] However, a significant problem is to sufficiently constrain the use of "free" undetermined variables, that is, to be sensitive to the danger that a fit may be obtained by attributing values needed to one or more unmeasured, unspecified, or possibly hidden variables involved.

De Alessi made a number of remarks on my use of the word personality and the fact that some aspects of X-efficiency theory may be nonobservables. But property rights theory depends on individuals' preferences and on utility maximization. It is obviously a "personality" who possesses nonobservable preferences and who does the nonobservable utility maximizing. Even property rights constraint changes may be unobservable. Is a change in feelings by a manager for his (or her) spouse who owns a significant share of a company observable?

De Alessi argues that under monopoly, costs will be higher because monopolies need more information about demand. However, this is not the same thing as the X-efficiency theory conclusion. The X-efficiency theory inference is not the *selling* costs, or market search costs, which are higher but the production costs which have nothing to do with finding out more about the demand function. Clearly, the conclusion is not only that costs are higher, but the specific nature of the costs. In Walter Primeaux's (1977) article, as well as some other work, the difference between the sheltered firm and the less sheltered one is in terms of production costs.

On the *F*-twist methodological issue, I believe that counterfactual postulates are unlikely to lead to correct *coherent* explanations. If the postulates cannot be tested, then we are forced to consider only the implications. But we would believe there is something wrong if we had a theory whose postulates were known to be counter to fact but which lead to correct predictions. This whole matter has been discussed from various viewpoints by a variety of authors (see Mark Blaug, 1980). I do not believe that the only purpose of theory is as an engine for prediction, nor do I see that we should look at any particular set of methodological views as imposing decisive constraints on our scientific procedures at this stage in our knowledge.[19]

De Alessi seems to object to the use of the neoclassical model as a standard for measuring inefficiency. But we need not compare an actual situation with a pie in the sky ideal, but rather with some particular performance such as best practice performance. The ideal may be superior to best practice. If anything, it seems to be a measurement problem. Careful procedures for studies of this sort have been worked out by econometricians who attempt to measure *frontier* production functions, as well as to measure deviations from the frontier.[20]

---

[17] Einstein's general theory of relativity would be just as valid if the measurement of the deflection of light by the sun during the 1919 eclipse had been carried out and known *before* the theory was developed rather than afterward. It is hard to see why De Alessi is concerned about my presumed *ex post* procedures. Since I am not an instrumentalist, or a single purpose theorist, I do not have to be only concerned with predictions.

[18] There is a sense in which some of the studies I referred to were *ex ante*. They were carried out by individuals that I did not know. They noted the inferences from X-efficiency theory and they happened to have evidence to process which would confront the inferences. In a few cases, the evidence came from a comparison of situations that came fairly close to results one would obtain from an experiment. These results could have been published whether or not they were in conformity with X-efficiency theory.

[19] For a discussion of instrumentalism and carefully worked out arguments against it, see Ernest Nagel (1961, pp. 129–52). The appeal to Occam's razor is surely besides the point, if as I believe I have shown, we indeed have different postulates and different theories.

[20] See the *Journal of Econometrics*, May 1980, especially the introductory paper by F. R. Forsund, C. A. K. Lovell, and P. Schmidt.

In general, we cannot understand the world well unless we distinguish between the inside and outside of firms, even if we do focus on individual behavior inside. The reason is simple, but perhaps not obvious. The behavior of the firm in the market requires a different sort of behavior of its members than that required inside. The firm, and the people who represent it *to* the market, may have to be aggressive maximizers; but inside the firm cooperation and nonmaximizing behavior comes closer to optimality. Inside what matters is behavior according to such values as commitment to the objectives of the firm and not to one's own objectives, loyalty to the firm, trust, commitment to a work ethic, and so on. As an economy develops firms get larger. The ratio of economic activities inside vs. outside the firm probably increases. As a result, efficiency (relative to capital and technical knowledge) may fall if the values significant for pursuing the firm's interests outside, in a relatively competitive market, become the motivations and values used inside.

## VI. Conclusions

This is mostly an unnecessary controversy. In part, De Alessi and I are concerned with different questions and different analytical ideas. In the present limited state of our knowledge, there is room for the development of more than one approach to understand intrafirm and related economic behavior. To my mind, X-efficiency theory is a research program. It is an open theory that represents work in process. Certainly there is room for further work, improvements, and learning.

In this paper the following basic differences have been stressed: 1) X-efficiency theory allows for degrees of inefficiency. Property rights theory argues that all decisions are efficient. 2) X-efficiency theory relaxes the maximization assumption and allows for nonmaximizing decisions when the sum of external and internal pressures is relatively low or beyond optimal pressure levels. Some types of nonmaximizing behavior help the organization to overcome free-rider incentives and the adoption of effort conventions. Property rights theory assumes that agents in all decision instances maximize utility subject to constraints. 3) X-efficiency theory separates exchange from intrafirm "activities" and asks how well nonmarket activities are organized. Property rights theory seems to tie everything to explicit or implicit exchange contracts. 4) The X-efficiency approach attempts to consider a variety of motivating forces within the firm, including environmental forces, and it accepts property rights as one such motivating force. 5) X-efficiency theory focuses on the interaction of all firm members (including both managers and employees) in their horizontal, vertical, and historical relationships. These interactions may be adversarial or cooperative both between and within groups, and they involve the possible responses of individuals to free-rider incentives. Property rights theory appears to ignore most of these basic intrafirm nonmarket interactions.

## REFERENCES

**Arrow, Kenneth J.,** *The Limits of Organization*, New York: W. W. Norton, 1974.

**Atkinson, J. W. and Birch, O.,** *Introduction to Motivation*, New York: Van Nostrand, 1978.

**Aumann, Robert J.,** "Survey of Repeated Games," *Essays in Game Theory*, Mannheim: Wissernschaftsverlag, 1981.

**Blaug, Mark,** *The Methodology of Economics*, Cambridge: Cambridge University Press, 1980.

**Cornell, Alexander H.,** *The Decision Makers' Handbook*, Englewood Cliffs: Prentice-Hall, 1980.

**De Alessi, Louis,** "Property Rights, Transaction Costs, and X-Efficiency: An Essay in Economic Theory," *American Economic Review*, March 1983, *73*, 64–81.

**Forsund, F. R., Lovell, C. A. K., and Schmidt, P.,** "A Survey of Frontier Production Functions and of their Relationship to Efficiency Measurement," *Journal of Econometrics*, May 1980, *13*, No. 1, 5–25.

**Jensen, Michael C. and Meckling, William H.,** "Rights and Production Functions: An Application to Labor-Managed Firms and Codetermination," *Journal of Business*, October 1979, *52*, 469–506.

**Leibenstein, H.,** *Beyond Economic Man*, Cambridge: Harvard University Press, 1976.

______, *General X-Efficiency Theory and Economic Development*, New York: Oxford University Press, 1978.

______, "Economic Decision Theory and Human Fertility Behavior," *Population and Development Review*, No. 3, 1981, *7*, 381–400.

______, "The Prisoners' Dilemma in the Invisible Hand: An Analysis of Intrafirm Productivity," *American Economic Review Proceedings*, May 1982, *72*, 92–97.

**Lewis, D.,** *Convention: A Philosophical Study*, Cambridge: Harvard University Press, 1969.

**Miles, Robert H.,** *Macro Organizational Behavior*, Santa Monica: Goodyear Publishing, 1980.

**Nagel, Ernest,** *The Structure of Science*, New York: Harcourt, Brace, 1961.

**Nelson, Richard R. and Winters, Sidney G.,** *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

**North, Douglass,** review of *The Rise and Decline of Nations* by Mancur Olson, *Science*, January 14, 1983, *219*, No. 4581, 163–64.

**Pratten, C. F.,** *Labor Productivity Differentials within International Companies*, Cambridge: Cambridge University Press, 1976.

**Primeaux, Walter J., Jr.,** "An Assessment of X-Efficiency Gained through Competition," *Review of Economics and Statistics*, February 1977, *59*, 105–08.

______, "An Assessment of the Effect of Competition on Advertising Intensity," *Economic Inquiry*, October 1981, *19*, 613–25.

**Rapoport, A.,** "Prisoner's Dilemma—Recollections," in his *A Game Theory as a Theory of Conflict Resolution*, Dordrecht: Reidel, 1974, 17–34.

**Rattner, Steven,** "A Tale of Two Ford Plants," *New York Times*, October 13, 1981.

**Schelling, T. S.,** *The Strategy of Conflict*, Oxford: Oxford University Press, 1960.

**Schotter, Andrew,** *Economic Theory of Social Institutions*, Cambridge: Cambridge University Press, 1981.

**Simon, H. A.,** "Rationality as Process and as Product of Thought," *American Economic Review Proceedings*, May 1978, *68*, 1–16.

**Smiley, Robert,** "Tender Offers, Transactions Costs and the Theory of the Firm," *Review of Economics and Statistics*, February 1976, *58*, 22–32.

**Ullmann-Margalit, E.,** *The Emergence of Norms*, New York: Oxford University Press, 1977.

**Williamson, Oliver E.,** *Corporate Control and Business Behavior*, Englewood Cliffs: Prentice-Hall, 1970.

# Property Rights and X-Efficiency: Reply

By LOUIS DE ALESSI*

Recently (1983), I assessed the relative merits of X-efficiency and of neoclassical economic theory generalized to include the constraints imposed by transaction costs and the structure of property rights. On both theoretical and empirical grounds, I concluded that such a generalization encompassed whatever limited explanatory power X-efficiency had to offer.

Harvey Leibenstein's comment (1983) sharpens his presentation of X-efficiency and, in my view, highlights the limitations of that construct.[1] Because the main points he raises were already addressed in my original article, further discussion would be repetitive. Leibenstein, however, seems to misinterpret several key aspects of my analysis. Accordingly, it is useful to restate these points to avoid misunderstanding by other readers and to facilitate assessment of the positions at issue.

Neoclassical economic theory, defined as the utility-maximization theory of the consumer and the wealth-maximization theory of the firm, assumed—inter alia—that all rights to the use of resources were fully allocated, privately held, and exchanged at zero information and transaction costs. By the late 1950's, empirical studies began to reveal some serious limitations of this framework and stimulated a variety of proposed reforms.

Different lines of research eventually led to the generalization of neoclassical theory to include the explicit consideration of transaction costs and the structure of property rights. This approach has turned out to be extremely fruitful. Not only has it reconciled the theory with broad classes of previously unexplained phenomena, but it has opened the way for important new applications. In particular, it has allowed extension of the utility-maximization hypothesis to *all* choices (i.e., by household, business, and government decision makers) subject to constraints. This advance has ended the dichotomy between the theory of consumer choice and the theory of the firm, and has established a more powerful framework for investigating the organization of economic activity.

Thus, for example, analysis of the shirking-information problem of team production is offering new insights into why firms exist. Indeed, it is providing a new perspective on vertical integration, occasioning a major reappraisal of antitrust policies; on the circumstances affecting the choice of business organization, including single proprietorships, partnerships and other profit-sharing arrangements, corporations, nonprofits, and government-owned firms; and on the economic consequences of alternative ownership arrangements. Both theory and evidence indicate that the choice of organizational form and the rules adopted within it depend crucially not only on transaction costs, but also on the formal legal system as well as on the work ethic, customs, conventions, and other components of a society's system of property rights.

Transaction costs and the system of property rights simply belong to the group of variables that determine the opportunity set faced by a utility-maximizing decision maker. It follows that "property rights," like transaction costs, is not a theory (it is a variable that enters economic theory); it is not a motivational force (other than in the sense that it affects the opportunity set); it does not imply that individuals act to fully maximize wealth given the limits of their property rights (its most distinguishing feature is the insistence upon the universal applicability of the utility-maximization hypothesis); it deals with *all* individual choices subject to constraints (including those made within business and government organizations as well as those that entail exchange); and it refers to the whole range of man-made institutional constraints (i.e., legal relationships, customs, ethics, and all conventions).

---

*Law and Economics Center, University of Miami, P.O. Box 248000, Coral Gables, FL 33124. I am indebted to Roger E. Meiners and Robert J. Staaf for helpful comments.

[1] Leibenstein's statement that I should have considered other alternative approaches, including Williamson's, is mystifying. First, I was not writing a review of the literature. Second, Williamson's work is a major keystone of the analysis I presented. My paper contained several statements to that effect plus repeated references to his writings.

The issue of efficiency is vexing. Following George Stigler (1976) and others, I defined efficiency as the solution to a constrained maximization problem. Different constraints yield different efficient equilibrium solutions, and theory may be used to explore the consequences of alternative constraints. Thus, one set of constraints may yield an efficient equilibrium solution entailing a negligible amount of shirking (for example, few errors, little sloppiness) and production of the measured output at costs approaching those of the ideal neoclassical equilibrium solution. A different set of constraints that lowers the cost of shirking would yield a new, efficient equilibrium solution with more shirking (for example, more errors and sloppiness) and higher production costs for the measured output. To view the latter solution as less efficient, however, is to ignore the utility of the shirking and the distribution of the gains and losses under the alternative sets of property rights. The history and present state of formal welfare economics suggest that these are shifting sands. Labeling some outcome inefficient, X or otherwise, simply does not solve the question of standards or provide any benchmark for predicting choices.

To clarify the issue further, consider Frank H. Knight's statement that

> There is a common misconception that it is possible to measure or discuss efficiency in purely physical terms. The first principles of physics or engineering science teach that this is not true, that the term efficiency involves the idea of value, and some measure of value as well. It is perhaps the most important principle of physical science that neither matter nor energy can be created or destroyed, that whatever goes into any process must come out in some form, and hence as a mere matter of physical quantity, the efficiency of all operations would equal one hundred per cent. [1951, pp. 9–10]

Leibenstein presumably would agree that if engine $A$ is technologically more efficient than engine $B$ (i.e., $A$ yields more thrust per unit of fuel), it does not follow that engine $A$ is economically more efficient (the cost of building $A$ could be sufficiently higher to offset the lower fuel costs). Similarly, however, if engine $A$ is economically more efficient in producing the measured output (i.e., in equilibrium it yields thrust at a lower unit cost), it does not follow that it is economically more efficient in a broader sense (for example, it could yield more air or noise pollution). Moreover, even if engine $A$ is more efficient in a broader sense, it does not follow that it will be chosen. Or, indeed, that it *should* be chosen, given the absence of a costless system for providing full compensation. As in the case of any other set of options, decision makers will maximize their utility (which includes their view of the public interest) subject to the applicable constraints.

It seems convenient at this point to address the issue of maximizing vs. nonmaximizing behavior. Generalized neoclassical theory implies that, under a variety of alternative constraints, utility-maximizing individuals will shirk or engage in other behavior that Leibenstein insists should be viewed as nonmaximizing. In the absence of any evidence to the contrary, however, and none was offered, an increase in the cost of such nonmaximizing behavior will induce at least some individuals to do less of it. Accordingly, the nonmaximizing label seems idiosyncratic and misleading.

Taking transaction costs and the system of property rights into account is providing new insights into a broad range of previously unexplained phenomena, including those addressed by X-efficiency. It does not follow, of course, that these advances exhaust the range of possible improvements in the theory, or that the theory may not in time be replaced by another, simpler and more powerful theory. In assessing the relative merits of the existing framework against alternative competing formulations, however, it is necessary to take into account the relevant state of the world—including transaction costs and the system of property rights—and to eliminate irrelevant statements by the rigorous application of Occam's razor.

At least some of the difficulty in coming to grips with X-efficiency results from Leibenstein's proclivity for redefining terms to suit his convenience, thereby setting up suitable straw men, and for seeking to refute hypotheses by musing about conceivable circumstances that would modify or deny them.[2] The meaning of any hypothesis, however, resides in the conceivable behavior that it denies. Regardless of the purposes to which any scientific theory might be put—prediction, explanation, guide to policy, and so on—ultimately its usefulness depends upon its ability to deal with real world phenomena. As a result, its truth or falsity ultimately must be established empirically (for example, see Ernest Nagel, 1963).[3]

The analysis presented in my earlier paper indicates that generalized neoclassical theory is superior to the X-efficiency construct on both theoretical and empirical grounds. Generalized neoclassical theory is simpler and yields a richer, broader range of implications, including all of those supposedly generated by X-efficiency, and receives broader, more rigorous empirical support. Although Leibenstein argues that the content of X-efficiency differs from the content of generalized neoclassical theory, he fails to provide any testable hypotheses or any empirical evidence capable of supporting that claim.[4]

---

[2] Leibenstein seems to be concerned with the rationality of the individual and how particular individuals might behave in particular circumstances. My concern, on the other hand, is with the rationality (logic) of the theory and its ability to explain real world phenomena.

[3] Leibenstein also footnotes Nagel (1961) approvingly.

[4] In its own terms, X-efficiency is X-inefficient.

## REFERENCES

**De Alessi, Louis,** "Property Rights, Transaction Costs, and X-Efficiency: An Essay in Economic Theory," *American Economic Review*, March 1983, *73*, 64–81.

**Knight, Frank H.,** *The Economic Organization*, New York: August M. Kelley, 1951.

**Leibenstein, Harvey,** "Property Rights and X-Efficiency: Comment," *American Economic Review*, September 1983, *73*, 831–42.

**Nagel, Ernest,** "Assumptions in Economic Theory," *American Economic Review Proceedings*, May 1963, *53*, 211–19.

______, *The Structure of Science*, New York: Harcourt, Brace, 1961.

**Stigler, George J.,** "The Xistence of X-Efficiency," *American Economic Review*, March 1976, *66*, 213–16.

# Preliminary Announcement of the Program

## NINETY-SIXTH ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

### San Francisco, California, December 27–30, 1983

**Tuesday, December 27, 1983**

10:00 A.M. EXECUTIVE COMMITTEE MEETING

**Wednesday, December 28, 1983**

8:00 A.M. HAS THE RESPONSE OF INFLATION TO MACRO POLICY CHANGED?*
*Presiding*: ROBERT E. HALL, Hoover Institution
*Papers*: JOHN B. TAYLOR, Princeton University
Macroeconomic Policy and Changes in the Cyclical Behavior of Wages
OLIVIER BLANCHARD, Harvard University
Has There Been a Change of Regime?
OTTO ECKSTEIN, Harvard University and Data Resources, Inc.
The Disinflation of 1980–83: How Did It Happen?
*Discussants*: STEVEN ENGLANDER, Federal Reserve Bank of New York
CORNELIS LOS, Federal Reserve Bank of New York

8:00 A.M. ECONOMIC EFFECTS OF THE RISE IN ENERGY PRICES: WHAT WE HAVE LEARNED IN TEN YEARS*
*Presiding*: ROBERT J. GORDON, Northwestern University
*Papers*: DALE W. JORGENSON, Harvard University
The Impact of Energy Prices on Capital and Labor Productivity: A Reassessment
JAMES L. SWEENEY, Stanford University
What We Have Learned About the Response of Energy Demand to Higher Energy Prices
ROBERT J. GORDON, Northwestern University
Implications of Supply Shocks for Macroeconomic Policy
*Discussants*: ERNST BERNDT, Massachusetts Institute of Technology
WILLIAM W. HOGAN, Harvard University
ALLAN H. MELTZER, Carnegie-Mellon University

8:00 A.M. RATIONALITY AND IRRATIONALITY IN FINANCIAL MARKETS
*Presiding*: RICHARD THALER, Cornell University
*Papers*: DANIEL KAHNEMAN, University of British Columbia, AND AMOS TVERSKY, Stanford University
Transparent and Opaque Preferences
THOMAS RUSSELL, University of Santa Clara, AND RICHARD THALER, Cornell University
The Relevance of Quasi Rationality in Competitive Markets
RICHARD J. ZECKHAUSER, Harvard University, AND VICTOR NIEDERHOFFER, NCZ Inc.
How Rational are the Participants in Financial Markets?
*Discussants*: KENNETH J. ARROW, Stanford University
WILLIAM J. BAUMOL, Princeton University and New York University

8:00 A.M. STRUCTURAL AND AGGREGATIVE METHODOLOGIES IN HISTORICAL PERSPECTIVE (Joint Session with the History of Economics Society)
*Presiding*: LAWRENCE A. BOLAND, Simon Fraser University
*Papers*: JOHN M. LETICHE, University of California-Berkeley
Linking Mainstream Macroeconomics, Microeconomics, and Keynesian Economics
DONALD E. MOGGRIDGE, University of Toronto
On Keynes's Method
LAWRENCE S. MOSS, Babson College
The "Ricardo Effect" Controversy in Light of Keynes
*Discussants*: LAWRENCE A. BOLAND, Simon Fraser University
DAVID LEVY, George Mason College, University of Virginia

8:00 A.M. STABILIZATION AND ADJUSTMENT POLICIES IN DEVELOPING COUNTRIES (Joint Session with the National Economic Association)
*Presiding*: DARIUS MANS, Federal Reserve Board
*Papers*: DARIUS MANS, Federal Reserve Board
Stabilization Programs and Economic Growth or: Whatever Happened to Economic Development?
RONALD JOHNSON, International Monetary Fund
The Response of Private International Capital Flows to Fund Programs
TETTEH KOFI, University of San Francisco
An Analysis of Recent Stabilization Policies in West Africa
GAVIN CHEN, U.S. Department of Commerce, AND ALI ABDI, International Monetary Fund
The Role of Hot Money During Economic Stabilization Programs with Special Reference to the Small, Open Economies of the Carribean
*Discussants*: ROBERT BROWN, Howard University
WILLIAM DARITY, JR., University of North Carolina-Chapel Hill

8:00 A.M. MACROECONOMIC ANALYSIS WITH DISEQUILIBRIUM MODELS: OPEN ECONOMIES AND INTERNATIONAL COMPARISONS
*Presiding*: RICHARD E. QUANDT, Princeton University
*Papers*: JOHN T. CUDDINGTON, Stanford University
A Fix-Price Trade Model with Perfect Capital Mobility: Fixed vs. Flexible Exchange Rates
HADI ESFAHANI, University of California-Berkeley
Macroeconomic Adjustment with Intermediate Imports, Fixed Prices, and Rationing
RICHARD F. KOSOBUD, University of Illinois-Chicago
Supply Shocks, Policy Responses, and Output Slowdowns: A Cross-Country Comparison with a Disequilibrium Model
*Discussants*: ELLEN MEADE, Princeton University
OLIVER LANDMANN, University of Basel

8:00 A.M. RECENT EMPIRICAL RESEARCH IN LABOR ECONOMICS
*Presiding*: GEORGE JOHNSON, University of Michigan
*Papers*: STEPHEN G. CECCHETTI, New York University
The Frequency of Union Wage Adjustment, 1958 to 1978
W. LEE HANSEN, University of Wisconsin-Madison, DAVID C. STAPLETON, Dartmouth College, AND DOUGLAS J. YOUNG, Montana State University
Forecasting the Market for New Ph.D. Economists: Evaluation and Update
WILLIAM T. ALPERT AND JOHN GUERARD, Lehigh University
Employment, Unemployment, and the Minimum Wage: A Causality Model
ARTHUR E. BLAKEMORE AND DENNIS L. HOFFMAN, Arizona State University
Tenure On-the-Job and Short-Run Productivity in Manufacturing
*Discussants*: ALAN GUSTMAN, Dartmouth College
JOHN PENCAVEL, Stanford University

8:00 A.M. THE RAPIDLY EXPANDING FOR-PROFIT HOSPITAL SECTOR (Joint Session with Health Economics Research Organization)
*Presiding*: SAMUEL A. MITCHELL, Federation of American Hospitals
*Papers*: DEBORAH A. FREUND, University of North Carolina, MARSHALL RUFFIN, Stanford University, AND RICHARD SHACHTMAN, University of North Carolina
Analysis of Length of Stay Differences Between Proprietary and Voluntary Hospitals
DAVID I. KASS AND THOMAS R. MCCARTHY, Federal Trade Commission
The Effect of Certificate of Need on For-Profit Hospitals
EDMUND BECKER, American Medical Association, AND FRANK SLOAN, Vanderbilt University
Hospital Ownership and Performance
*Discussants*: HAROLD TING, National Medical Enterprises
JON GABEL, National Center for Health Services Research, DHHS
RICHARD ERNST, University of Southern California

8:00 A.M. THE "NEW" LAW AND ECONOMICS
*Presiding*: ANN DRYDEN WITTE, University of North Carolina
*Papers*: DAVID D. HADDOCK AND CHRISTOPHER CURRAN, Emory University Law and Economics Center
Negligence, Comparative Negligence, and Double Liability: The Empirical Notion of Efficiency in Modeling Legal Rules of Liability

SUSAN ROSE-ACKERMAN, Columbia University School of Law
Inalienability and Compensation: Another View of the Cathedral
WERNER Z. HIRSCH, University of California-Los Angeles
Economic Effects on Indigent Black Tenants of Landlord Tenant Laws
JUDITH A. LACHMAN, University of Wisconsin-Madison, School of Law
Freedom of Speech: Theory and Historical Analysis of Case Law
*Discussants*: HENRY MANNE, Emory University
ROBERT COOTER, University of California-Berkeley
DANIEL RUBINFELD, University of Michigan
ELI NOAM, Columbia University, AND LEWIS A. KORNHAUSER, New York University

10:15 A.M. INTERNATIONAL DIFFERENCES IN WAGE AND PRICE BEHAVIOR*
*Presiding*: ROBERT J. GORDON, Northwestern University
*Papers*: BERT G. HICKMAN, Stanford University, AND LAWRENCE R. KLEIN, University of Pennsylvania
Wage-Price Behavior in the National Models of Project LINK
GEORGE A. KAHN, Federal Reserve Bank of Kansas City
Nominal and Real Wage Stickiness: Evidence from Six Countries
CHARLES L. SCHULTZE, The Brookings Institution
The Cyclical Responsiveness of Inflation: Implications of Cross-Country and Cross-Temporal Differences
*Discussants*: OLIVIER BLANCHARD, Harvard University
RICHARD T. FROYEN, University of North Carolina
JEFFREY D. SACHS, Harvard University

10:15 A.M. THE APPROPRIATE RESPONSE TO TRADE BARRIERS AND "UNFAIR" TRADE PRACTICES IN OTHER COUNTRIES*
*Presiding*: ROBERT E. BALDWIN, University of Wisconsin-Madison
*Papers*: ROBERT E. BALDWIN AND T. SCOTT THOMPSON, University of Wisconsin-Madison
Responding to "Unfair" Trade Practices: An Economic Viewpoint
JOHN JACKSON, University of Michigan
The Lawyer's Perspective: The Case for Attention to Procedures
STEPHEN D. KRASNER AND JUDITH GOLDSTEIN, Stanford University
Trade Barriers and Unfair Trade Practices: The Case for a Differential Response
*Discussants*: JOHN CUDDINGTON, Stanford University
JOHN ZYSMAN, University of California-Berkeley

10:15 A.M. WHAT SURVIVES OF THE RATIONAL EXPECTATIONS REVOLUTION?*
*Presiding*: STANLEY FISCHER, Massachusetts Institute of Technology
*Papers*: ROBERT J. BARRO, University of Chicago
What Survives of the Rational Expectations Revolution?
STEPHEN LEROY, University of California-Santa Barbara
Efficiency and the Variability of Asset Prices
JULIO ROTEMBERG, Massachusetts Institute of Technology
On the Statistical Failures of Rational Expectations Macroeconomic Models
*Discussants*: WILLIAM NORDHAUS, Yale University
ROBERT MERTON, Massachusetts Institute of Technology
KENNETH SINGLETON, Carnegie-Mellon University

10:15 A.M. IS THE UNITED STATES UNDERINVESTING IN PUBLIC CAPITAL?*
*Presiding*: EDWARD M. GRAMLICH, University of Michigan
*Papers*: CHARLES HULTEN AND GEORGE PETERSON, The Urban Institute
The Growth, or Lack of Growth, in the Public Capital Stock
MARTIN N. BAILY, The Brookings Institution
Does Public Capital Accumulation Affect Private Productivity Growth: A Look at the Transportation Industry
JAMES JONDROW AND ROBERT LEVY, Center for Naval Analyses
The Displacement of Local Spending for Pollution Control by Federal Construction Grants
*Discussants*: MICHAEL BOSKIN, Stanford University
LAURENCE KOTLIKOFF, Yale University
ROBERT REISCHAUER, The Urban Institute

10:15 A.M. ECONOMIC EDUCATION*
*Presiding*: W. LEE HANSEN, University of Wisconsin-Madison
*Papers*: GEORGE LELAND BACH, Stanford University, AND ALLEN C. KELLEY, Duke University
Improving the Teaching of Economics: Achievements and Aspirations
JOHN J. SIEGFRIED AND JENNIE E. RAYMOND, Vanderbilt University
A Profile of Students Majoring in Economics
*Discussants*: EDWIN S. MILLS, Princeton University
ALICE M. RIVLIN, The Brookings Institution
NATHAN ROSENBERG, Stanford University

10:15 A.M. PSYCHOLOGICAL AND SOCIOLOGICAL FOUNDATIONS OF ECONOMIC BEHAVIOR*
*Presiding*: GARY BECKER, University of Chicago
*Papers*: GEORGE A. AKERLOF, University of California-Berkeley
Why Macroeconomics Needs Sociology
JAMES S. COLEMAN, University of Chicago
Social Systems and Economic Analysis
ALBERT O. HIRSCHMAN, Institute for Advanced Study
Five Simple Ways of Complicating the Categories of Economic Discourse
*Discussants*: JOSEPH E. STIGLITZ, Princeton University
KENNETH J. ARROW, Stanford University
A. MICHAEL SPENCE, Harvard University

10:15 A.M. INFORMATION, ORGANIZATIONAL STRUCTURE AND MARKET PERFORMANCE
*Presiding*: JOHN G. RILEY, University of California-Los Angeles
*Papers*: PETER REUTER, The Rand Corporation
The Organization of Illegal Markets
SCOTT E. MASTEN, University of Virginia
Transactions Costs, Institutional Choice, and the Theory of the Firm
JANET KIHOLM SMITH AND STEVEN R. COX, Arizona State University
The Pricing of Legal Services: A Market Solution to the Problem of Bilateral Opportunism
KEITH J. CROCKER AND ARTHUR SNOW, University of Virginia
The Efficiency of Competitive Equilibria in Insurance Markets with Asymmetric Information
*Discussants*: BENJAMIN KLEIN, University of California-Los Angeles
KEITH LEFFLER, University of Washington
CHARLES WILSON, New York University
JEFFREY PERLOFF, University of California-Berkeley

10:15 A.M. NEW DEVELOPMENTS IN THE ANALYSIS OF PRODUCTION
*Presiding*: LAURITS CHRISTENSEN, University of Wisconsin-Madison
*Papers*: G. THOMAS SAV, West Georgia College
Micro Estimates of Energy-Capital Complementarity Using Engineering Production Functions
RAYMOND KOPP AND MICHAEL HAZILLA, Resources for the Future, Inc.
Testing for the Existence of a Consistent Aggregate Index of Capital Inputs: An Application of Approximation Analysis to the Multisectoral U.S. Production Technology
ROGER BETANCOURT, University of Maryland-College Park
A Generalization of Modern Production Theory
*Discussants*: ERNST BERNDT, Massachusetts Institute of Technology
DOUGLAS CAVES, University of Wisconsin-Madison
MICHAEL DENNY, University of Toronto

10:15 A.M. ECONOMIC SAFETY NETS
*Presiding*: ROBERT J. LAMPMAN, University of Wisconsin-Madison
*Papers*: TIMOTHY M. SMEEDING, University of Utah
Work Related Benefits—Their Role in the Safety Net
LEE BAWDEN AND JOHN PALMER, The Urban Institute
Reaganomics and the Social Safety Net
MARILYN MOON AND PATRICIA RUGGLES, Congressional Budget Office
Reaganomics and the Elderly: The 1983 Social Security Amendments and Other Changes
*Discussants*: SYLVESTER SCHEIBER, Employee Benefit Research Institute
SHELDON DANZIGER, University of Wisconsin-Madison
RICHARD V. BURKHAUSER, Vanderbilt University

10:15 A.M. INDUSTRIAL POLICY: EAST, WEST, AND SOUTH (Joint Session with the Association for Comparative Economic Studies)
*Presiding*: RICHARD R. NELSON, Yale University
*Papers*: THOMAS A. PUGEL, New York University
Japan
MERTON J. PECK, Yale University
United States
J. MICHAEL MONTIAS, Yale University, AND JOSEF BRADA, Arizona State University
Eastern Europe
(To be announced)
Developing Countries
*Discussants*: RICHARD R. NELSON, Yale University
MERTON J. PECK, Yale University

10:15 A.M. DEVELOPMENT POLICIES: A REASSESSMENT
*Presiding*: JAGDISH BAGHWATI, Columbia University
*Papers*: PAUL STREETEN, Boston University
Basic Needs: A Reassessment
BELA BALASSA, The Johns Hopkins University and World Bank
Adjustment Policies: A Reassessment
IRMA ADELMAN, University of California-Berkeley
Development Strategies: A Reassessment
*Discussants*: ALBERT FISHLOW, University of California-Berkeley
CARLOS-DIAZ ALEJANDRO, Columbia University

10:15 A.M. POLITICAL ECONOMY OF STATE GOVERNMENT DECISION MAKING, I: INSTITUTIONS AND PROCESSES
*Presiding*: ROGER BOLTON, Williams College
*Papers*: JOSEPH SENECA, Rutgers University and New Jersey Economic Policy Council
State Economic Advising: Issues, Methods, and Lessons
JOHN BRANDL, University of Minnesota and Minnesota House of Representatives
Distilling Frenzy from Academic Scribblers: On Being a Legislator
CHARLES RENFRO, Chase Econometrics
Through a Glass Darkly: Regional Governments and their Economic Advisers
*Discussants*: LARRY KIMBELL, University of California-Los Angeles
HELEN LADD, Harvard University

10:15 A.M. RACE, CRIME, AND EMPLOYMENT (Joint Session with the National Economic Association)
*Presiding*: SAMUEL L. MYERS, JR., University of Pittsburgh
*Papers*: LLAD PHILLIPS AND HAROLD ZOTEY, University of California-Santa Barbara
Female Employment, Race, and Crime
IRVING PILIAVIN AND REBECCA MAYNARD, University of Wisconsin-Madison
Work and Crime: Evidence from the Supported Work Experiment
DURAN BELL AND KEVIN LANG, University of California-Irvine
An Economic Model of Disposition of Juvenile Offenders
JOSEPH COOPER, Harvard University
Unemployment and Crime: Evidence from the NBER Young Men Employment Study
*Discussants*: ANN D. WITTE, University of North Carolina-Chapel Hill
STEVEN A. HOENACK, University of Minnesota
CHARLES Z. WILSON, ZZ and Associates

10:15 A.M. THE CHANGING EFFECT OF INFLATION ON FINANCIAL RETURNS (Joint Session with the American Finance Association)
*Presiding*: PATRIC H. HENDERSHOTT, Ohio State University and National Bureau of Economic Research
*Papers*: RICHARD H. CLARIDA AND BENJAMIN M. FRIEDMAN, Harvard University and National Bureau of Economic Research
Why Have Short-Term Real Interest Rates Been So High?
PATRIC H. HENDERSHOTT, Ohio State University and National Bureau of Economic Research
Expected Inflation, Real Debt Rates, and Equity Returns

JOHN HUIZINGA AND FREDERIC S. MISHKIN, University of Chicago and National Bureau of Economic Research
The Measurement of Short-Term Real Rates on Assets with Different Risk Characteristics
*Discussants*: ROGER CRAINE, University of California-Berkeley
WARD S. CURRAN, Trinity College
DONALD MULLINEAUX, Federal Reserve Bank of Philadelphia

10:15 A.M. MANAGING ENDOWMENTS IN TIME OF INFLATION AND DISINFLATION (Joint Session with the Association for the Study of the Grants Economy)
*Presiding*: KENNETH E. BOULDING, University of Colorado
*Papers*: ROBERT S. MAIN, Butler University
Management of Endowments under Inflation and Disinflation: Alternative Models
WILLIAM S. VICKREY, Columbia University
On the Fallacies Leading to Endowment Mismanagement
THOMAS F. WILSON, Wilson Associates
Certain Specific Features of Endowment Financing
*Discussants*: STEPHEN H. LONG, Syracuse University
BURTON A. WEISBROD, University of Wisconsin-Madison

12:30 P.M. JOINT LUNCHEON WITH THE AMERICAN ECONOMIC ASSOCIATION AND THE AMERICAN FINANCE ASSOCIATION
*Presiding*: CHARLES L. SCHULTZE, The Brookings Institution, AND JAMES C. VAN HORNE, Stanford University
*Speaker*: PAUL VOLCKER, Chairman, Board of Governors, Federal Reserve System

2:15 P.M. ECONOMIC THEORY AND MANAGEMENT EDUCATION*
*Presiding*: DAVID J. TEECE, University of California-Berkeley
*Papers*: DAVID J. TEECE, University of California-Berkeley, AND SIDNEY WINTER, Yale University
The Limits of Neoclassical Theory in Management Education
ROBERT G. HARRIS, University of California-Berkeley
Management Miseducation: Consequences for the Wealth of the Nation
RICHARD CAVES, Harvard University
Economic Analysis and the Quest for Strategic Advantage
*Discussants*: WILLIAM ABERNATHY, Harvard University
M. THERESE FLAHERTY, Harvard University
DAVID KREPS, Harvard University
F. M. SCHERER, Swarthmore College

2:15 P.M. ANTITRUST POLICY AND STRATEGIC BEHAVIOR*
*Presiding*: LAWRENCE J. WHITE, New York University
*Papers*: JOHN C. HILKE AND PHILLIP B. NELSON, Federal Trade Commission
Noisy Advertising and the Predation Rule in Antitrust Analysis
WILLIAM S. COMANOR AND H. E. FRECH III, University of California-Santa Barbara
Exclusive Dealing and Strategic Behavior
ROBERT E. DANSBY, Bell Laboratories
Commodity Bundling and Injunctive Pricing Restraints
*Discussants*: STEVEN C. SALOP, Georgetown University Law Center
RICHARD J. GILBERT, University of California-Berkeley
LEONARD WAVERMAN, Institute for Policy Analysis, University of Toronto

2:15 P.M. MONETARISM: LESSONS FROM THE POST-1979 EXPERIMENT*
*Presiding*: HENRY C. WALLICH, Board of Governors of the Federal Reserve System
*Papers*: BENJAMIN M. FRIEDMAN, Harvard University
How Useful Were Monetary Aggregate Targets During the Post-1979 Experiment?
BENNETT T. MCCALLUM, Carnegie-Mellon University
Monetarist Principles in the Light of Recent Experience
JAMES L. PIERCE, University of California-Berkeley
Did Financial Innovation Hurt the Great Monetarist Experiment?
*Discussants*: ALAN S. BLINDER, Princeton University
MILTON FRIEDMAN, Hoover Institution
ALLAN H. MELTZER, Carnegie-Mellon University

2:15 P.M. Alternative Views of the Productivity Puzzle
*Presiding*: Martin N. Baily, The Brookings Institution
*Papers*: Edward N. Wolff, New York University
Industrial Composition, Interindustry Effects, and the U.S. Productivity Slowdown
Andrew Sharpe, Department of Finance, Canada
A Comparative Analysis of the Canadian and American Productivity Slowdowns
Edward M. Miller, Rice University
Obsolescence and the Measurement of Capital
Joel D. Scheraga, Rutgers University
The Heterodox View of Growth Accounting
*Discussants*: Richard Nelson, Yale University
Gregory Jump, University of Toronto
Michael Arnold, Institute of Regional and Urban Studies

2:15 P.M. Economic Implications of Changing Trade Structures: Asian Countries
*Presiding*: Robert E. Baldwin, University of Wisconsin-Madison
*Papers*: Richard Hooley, University of Pittsburgh
The Role of Industrial Productivity Growth in Export Expansion: The Case of the Philippines and Malaysia
Sung Y. Kwack, Howard University
Developments in and Prospects for External Debt Position and Burden of Developing Countries: The Case of Korea
Frank S. T. Hsiao, University of Colorado-Boulder, and Mei-chu Wang Hsiao, University of Colorado-Denver
What Makes Asian *NIC*s Tic? An Analysis of Changes in Trade Structure
*Discussants*: John M. Letiche, University of California-Berkeley
Kanta Marwah, Carleton University
Woo Bong Lee, Bloomsburg State College

2:15 P.M. Impacts of Changes in Military Expenditures and Foreign Assistance
*Presiding*: Kenneth Boulding, Swarthmore College
*Papers*: Lawrence R. Klein, University of Pennsylvania
Foreign Assistance from Reduced Military Expenditures: Implications for Growth of Developing Nations
Martin C. McGuire, University of Maryland
Foreign Assistance, Investment and Defense: A Methodological Study with an Application to Israel, 1960–79
*Discussants*: Walter Isard, Cornell University
Michael Intriligator, University of California-Los Angeles

2:15 P.M. Political Economy of State Government Decision Making, II: Scope for Discretion and Innovation
*Presiding*: William Oakland, Tulane University
*Papers*: Roger Bolton, Williams College
A State's Economic and Fiscal Structure as a Diversified Portfolio
William Whitesell, Franklin and Marshall College
Opportunities for Innovation in State Banking Regulation
Charles McLure, Hoover Institution
The Do's and Don't's of State Tax Policies
*Discussants*: Neal Soss, Federal Reserve Bank of New York and Federal Reserve Board
Karl Case, Wellesley College

2:15 P.M. The Price Effects of Structural Changes in the Regulated Industries (Joint Session with the Transportation and Public Utilities Group)
*Presiding*: Jay A. Smith, Jr., University of North Florida
*Papers*: Allan D. Schuster, University of Texas-Austin
Expected Structural Changes in the Motor Carrier Industry: Impact on Transportation Costs and Pricing Strategies
Laurence T. Phillips, U.S. Department of Transportation, and Reuben Kyle III, Middle Tennessee State University
Airline Deregulation: Did Economists Promise Too Much or Too Little?
Walter G. Bolter, Bolter and Nilsson Consultants
The Appropriateness of Regulation as a Means of Implementing Social Goals in Future Telecommunications Markets
*Discussants*: Gayton Germane, Stanford University
Joseph Gillan, Illinois Commerce Commission

2:15 P.M. DEMAND FOR MENTAL HEALTH SERVICES (Joint Session with Health Economics Research Organization)
*Presiding*: DONALD E. YETT, University of Southern California
*Papers*: CONSTANCE M. HORGAN, National Center for Health Services Research, DHHS
The Demand for Ambulatory Mental Health Services
RICHARD M. SCHEFFLER AND CAROLYN ADAMS WATTS, University of California-Berkeley
Changes in the Demand for Mental Health Services Under Blue Cross/Blue Shield
ALVIN K. KLEVORICK, Yale University, AND THOMAS G. MCGUIRE, Boston University
Regulation, Insurance, and Demand and Supply for Mental Health Services
*Discussants*: RICHARD G. FRANK, University of Pittsburgh
WILLARD MANNING, The Rand Corporation
CHARLES P. HALL, Temple University

2:15 P.M. THE ECONOMICS OF CATASTROPHIC RISK (Joint Session with the Association of Environmental and Resource Economists)
*Presiding*: MAUREEN CROPPER, University of Maryland
*Papers*: TALBOT PAGE, California Institute of Technology
A Groye-Like Mechanism in Risk Assessment
PAUL SLOVIC, BARUCH FISCHOFF, AND SARAH LICHTENSTEIN, Decision Research, Inc.
Forecasting the Economic Impact of Catastrophic Accidents: Psychological Perspectives
HERMAN B. LEONARD AND RICHARD J. ZECKHAUSER, Harvard University
Public Provision of Insurance
*Discussants*: HOWARD KUNREUTHER, University of Pennsylvania
TRACY LEWIS, University of British Columbia
DAVID GRETHER, California Institute of Technology

2:15 P.M. ENERGY DEMAND AND SUPPLY: RECENT RESEARCH RESULTS FOR EUROPE AND THE UNITED STATES
*Presiding*: EDWARD W. ERICKSON, North Carolina State University
*Papers*: ROY BOYD AND CAROL DAHL, University of Wisconsin-Milwaukee
Impact of the Soviet Gas Pipeline on Western Europe's Energy Cost
CHRISTOPHER GARBACZ, University of Missouri-Rolla
Residential Energy Demand: A National Micro-Based Model
JAMES B. KURISH AND MICHAEL J. PIETTE, University of Hartford
The Impact of the Federal Residential Energy Conservation Tax Credit on Individual Household Energy Conservation Expenditures
*Discussants*: JAMES SWEENEY, Stanford University
JAMES CHALFANT, University of California-Berkeley
JAMES PLUMMER, Q.E.D. Research Inc.

2:15 P.M. EXCHANGE RATE DETERMINATION IN ALTERNATIVE ECONOMIC ENVIRONMENTS
*Presiding*: JOHN H. MAKIN, University of Washington
*Papers*: CHARLES LACIVITA, University of North Carolina-Greensboro
Exchange Rates in General Equilibrium
STEPHEN A. MEYER, University of Pennsylvania and Federal Reserve Bank of Philadelphia
Financial Markets and External Exchange Rate Shocks with a Basket Peg
MARK M. PITT, University of Minnesota
Smuggling and the Black Market for Foreign Exchange
*Discussants*: KENNETH ROGOFF, Federal Reserve Board
DOUGLAS D. PURVIS, Queens University-Kingston
ROBERT HODRICK, Northwestern University

4:30 P.M. ECONOMIC OUTLOOK AND ECONOMIC POLICY
*Presiding*: JUANITA KREPS, Duke University
*Panel*: ALAN GREENSPAN, Townsend-Greenspan & Company
ALICE RIVLIN, The Brookings Institution
GARDNER ACKLEY, University of Michigan

8:00 P.M. RICHARD T. ELY LECTURE*
*Presiding*: CHARLES L. SCHULTZE, The Brookings Institution
*Papers*: THOMAS C. SCHELLING, Harvard University
Self-Command in Practice, in Policy, and in a Theory of Rational Choice

**Thursday, December 29, 1983**

8:00 A.M. NEW EVIDENCE ON THE DEMAND FOR SAFETY AND THE BEHAVIORAL RESPONSE TO SAFETY REGULATION*
*Presiding*: LAURENCE C. ROSENBERG, National Science Foundation
*Papers*: CLIFFORD WINSTON, Massachusetts Institute of Technology
Consumer Demand for Automobile Safety Regulation: A Hedonic Analysis
WILLIAM T. DICKENS, University of California-Berkeley
Differences in Risk Premia in Union and Nonunion Wages and the Case for Occupational Safety Regulation
W. KIP VISCUSI, Duke University
Behavioral Response to Safety Regulations: The Effect of Child Resistant Packaging on Aspirin Ingestion
ROBERT W. CRANDALL AND JOHN GRAHAM, The Brookings Institution
Automobile Safety Regulation and Offsetting Driver Behavior: Some New Evidence
*Discussant*: GLENN BLOMQUIST, University of Kentucky

8:00 A.M. INFORMATION AND MACROECONOMICS*
*Presiding*: EDMUND S. PHELPS, Columbia University
*Papers*: BRUCE GREENWALD, Harvard University and Bell Laboratories, JOSEPH E. STIGLITZ, Princeton University and Bell Laboratories, AND ANDREW WEISS, Columbia University and Bell Laboratories
Informational Imperfections and Macroeconomic Fluctuations
CARL SHAPIRO, Princeton University
Price Signaling, Price Rigidity and Macroeconomic Adjustments
JANET YELLEN, University of California-Berkeley
Uncertainty in Macroeconomics: A Survey
*Discussants*: GEORGE AKERLOF, University of California-Berkeley
ROBERT E. HALL, Stanford University

8:00 A.M. THE POLITICAL ECONOMY OF BUDGET DEFICITS*
*Presiding*: JOSEPH PECHMAN, The Brookings Institution
*Papers*: ALICE RIVLIN, The Brookings Institution
Reform of the Budget Process
ROBERT EISNER, Northwestern University
Which Budget Deficit? Some Issues of Measurement and their Implications
ALAN BLINDER AND DOUGLAS HOLTZ-EAKIN, Princeton University
Public Opinion Toward Deficits
*Discussants*: MURRAY WEIDENBAUM, Washington University
LAWRENCE SUMMERS, Harvard University
GEORGE VON FURSTENBERG, University of Indiana

8:00 A.M. THE THEORY OF CONTRACTS AND AGENCY: IMPLICATIONS FOR THE BEHAVIOR OF WAGES*
*Presiding*: MARTIN N. BAILY, The Brookings Institution
*Papers*: BENJAMIN KLEIN, University of California-Los Angeles
An Economic Theory of Rigid Wages
EDWARD LAZEAR, University of Chicago
Incentives and Wage Rigidity
ROBERT J. FLANAGAN, Stanford University
Implicit Contracts, Explicit Contracts, and Wages
*Discussants*: MARTIN N. BAILY, The Brookings Institution
JOHN HALTIWANGER, University of California-Los Angeles
VICTOR GOLDBERG, Northwestern University

8:00 A.M. IMMIGRATION AND EMPLOYMENT (Joint Session with the Industrial Relations Research Association)
*Presiding*: F. RAY MARSHALL, University of Texas
*Papers*: GREGORY DEFREITAS, Barnard College, Columbia University, AND ADRIANA MARSHALL, New York University
Immigration, Employment, and Wage Growth in U.S. Metropolitan Areas: 1970–80
MARCIA FREEDMAN, Columbia University
Employment Experience of Immigrants in the New York City Labor Market
PHILIP L. MARTIN AND RICHARD MINES, University of California-Davis
Immigration and the California Farm Labor Force
*Discussants*: VERNON BRIGGS, New York State School of Industrial and Labor Relations, Cornell University
RUSS WILLIAMS, Agricultural Producers, California
MUZAFFAR CHISHTI, International Ladies Garment Workers Union

8:00 A.M. ECONOMIC INEQUALITY: MEASUREMENT AND POLICY (Joint Session with the Econometric Society)
*Presiding*: GEORGE F. RHODES, JR., Colorado State University
*Papers*: ROBERT L. BASMANN, Texas A&M University
A Multivariate Approach to Modeling Comprehensive Economic Inequality
JAMES FOSTER, Purdue University
On Economic Poverty
DALE W. JORGENSON, Harvard University, AND DANIEL T. SLESNICK, University of Texas-Austin
Inequality in the Distributions of Individual Welfare
*Discussants*: GLENN LOURY, Harvard University
R. ROBERT RUSSELL, New York University
THOMAS M. STOKER, Massachusetts Institute of Technology

8:00 A.M. MEN, WOMEN, AND SOCIAL SECURITY
*Presiding*: ALICIA MUNNELL, Federal Reserve Bank of Boston
*Papers*: PAUL CULLINAN AND PATRICIA RUGGLES, Congressional Budget Office
Relative Benefits Received by Men and Women under Social Security: Current Laws and Two Options
BARBARA LIBBY AND ELIZABETH DURAN, Niagara University
Women Workers' Subsidization of the Social Security Program
MARJORIE HONIG, Hunter College
Labor Supply and Retirement Behavior of Older Women
*Discussants*: LAURENCE KOTLIKOFF, Yale University
SUSAN WACHTER, University of Pennsylvania
SOPHIE KORCZYK, Employee Benefit Research Institute

8:00 A.M. TOPICS IN URBAN AND REGIONAL ECONOMICS
*Presiding*: A. THOMAS KING, Federal Home Loan Mortgage Corporation
*Papers*: DAVID E. MILLS, Vanderbilt University
Indivisibilities and Land Development Timing: The Shopping Center Problem
JANICE FANNING MADDEN, University of Pennsylvania
Urban Wage Gradients: Empirical Evidence
CURTIS C. HARRIS AND MEHRZAD NADJI, University of Maryland
Spatial Modeling of Production Capacities
JOSEPH A. ZIEGLER, University of Arkansas
A Disequilibrium Model of Regional Employment Changes
*Discussants*: DONALD C. SHOUP, University of California-Los Angeles
RANDALL W. EBERTS, University of Oregon

8:00 A.M. INTERNATIONAL BANKING (Joint Session with the American Finance Association)
*Presiding*: ANDREW F. BRIMMER, Brimmer and Company, Inc.
*Papers*: LELAND S. PRUSSIA, Bank of America
Innovations and Structural Changes in International Banking
HENRY C. WALLICH, Board of Governors of the Federal Reserve System
International Commercial Banking from a Central Bank Viewpoint
ANDREW F. BRIMMER, Brimmer and Company, Inc.
International Banking: A Capital Markets Perspective
*Discussants*: JACK M. GUTTENTAG, University of Pennsylvania
HELEN B. JUNZ, International Monetary Fund

8:00 A.M. ECONOMIC AND BUSINESS OUTLOOK (Joint Session with the National Association of Business Economists)
*Presiding*: DONALD H. STRASZHEIM, Wharton Econometric Forecasting Associates
*Papers*: KEVIN R. KELLY, U.S. Bancorp
Consumer Spending
LYNN O. MICHAELIS, Weyerhaeuser Company
The Housing Outlook
DUANE A. PAUL, Bank of America
Plant and Equipment Spending
*Discussants*: DOUGLAS B. ATKINSON, PPG Industries, Inc.
JAMES F. SMITH, Union Carbide Corporation

8:00 A.M. MARKET IMPACTS OF HEALTH INSURANCE PAYMENTS MECHANISMS (Joint Session with the Health Economics Research Organization)
*Presiding*: BRYAN LUCE, Health Care Financing Administration, DHHS
*Papers*: AGNES RUPP, DONALD STEINWACHS, AND DAVID SALKEVER, The Johns Hopkins University
The Effect of Hospital Payment Methods on the Pattern and Cost of Mental Care
DONALD E. YETT, University of Southern California, WILLIAM DER, TransAmerica Corporation, RICHARD L. ERNST, University of Southern California, AND JOEL W. HAY, University of Connecticut Health Center
Fee Screen Reimbursement and Physician Fee Inflation
RONALD J. VOGEL, University of Arizona
On Health Cost Containment: An Analysis of Structural Incentives in General, and of the Arizona Health Care Cost Containment System in Particular
*Discussants*: J. MICHAEL FITZMAURICE, Health Care Financing Administration, DHHS
KATHLEEN ADAMS, American Medical Association
HAROLD S. LUFT, University of California-San Francisco

8:00 A.M. LIMITS TO GROWTH: WHAT HAVE WE LEARNED? (Joint Session with the Association of Environmental and Resource Economists, and the American Association for the Advancement of Science)
*Presiding*: ROBERT S. PINDYCK, Massachusetts Institute of Technology
*Panel*: WILLIAM NORDHAUS, Yale University
THOMAS C. SCHELLING, Harvard University
HERMAN DALY, Louisiana State University
JULIAN SIMON, University of Illinois

8:00 A.M. MODELING RESOURCE SUPPLY (Joint Session with the Association of Environmental and Resource Economists)
*Presiding*: ANTHONY C. FISHER, University of California-Berkeley
*Papers*: BRIAN J. SKINNER, Yale University
Estimation and Uses of Copper Resources: Geological Basis
TJALLING C. KOOPMANS, Yale University
Estimation and Uses of Copper Resources: Long-Range Projections of Production and Uses
MARTIN B. ZIMMERMAN, University of Michigan
Models of Mineral Supply and Measures of Scarcity
*Discussants*: DEVERELE HARRIS, University of Arizona
(Others to be announced)

10:15 A.M. A DECADE OF GREATER VARIABILITY IN EXCHANGE RATES: AN ASSESSMENT*
*Presiding*: RALPH C. BRYANT, The Brookings Institution
*Papers*: MARINA V.N. WHITMAN, General Motors Corporation
A Perspective on the Behavior of Firms and Financial Intermediaries
JEFFREY R. SHAFER, Federal Reserve Bank of New York
Macroeconomic Policies and Domestic Economic Performance
VAL KOROMZAY, JOHN LLEWELLYN, AND STEPHEN POTTER, Organization for Economic Cooperation and Development
Exchange Rates and Policy Choices: Lessons from Interdependence in a Multilateral Perspective
*Discussants*: PAUL ARMINGTON, Armington Wolford and Associates
JACOB A. FRENKEL, University of Chicago
ANDREW D. CROCKETT, International Monetary Fund

10:15 A.M. TAX POLICY: A FURTHER LOOK AT SUPPLY-SIDE EFFECTS*
*Presiding*: RUDOLPH G. PENNER, The American Enterprise Institute
*Papers*: CHARLES R. HULTEN, The Urban Institute
Tax Policy and the Investment Decision
JERRY A. HAUSMAN, Massachusetts Institute of Technology
Tax Policy and Labor Supply
LAWRENCE H. SUMMERS, Massachusetts Institute of Technology
Tax Policy and the Savings Decision
*Discussants*: ROGER GORDON, Bell Laboratories
MARK KILLINGSWORTH, Rutgers University
PHILIP HOWREY, University of Michigan

10:15 A.M. REFLECTIONS ON THE CURRENT STATE OF MACRO ECONOMIC THEORY*
*Presiding*: PAUL MCCRACKEN, University of Michigan
*Papers*: GEORGE PERRY, The Brookings Institution
(Title to be announced)
THOMAS SARGENT, University of Minnesota
(Title to be announced)
*Discussants*: ROBERT J. BARRO, University of Chicago
ALAN BLINDER, Princeton University
ROBERT E. HALL, Hoover Institution
WILLIAM NORDHAUS, Yale University

10:15 A.M. THEORETICAL ASPECTS OF PUBLIC EXPENDITURES AND PUBLIC CHOICE
*Presiding*: BRYAN ELLICKSON, University of California-Los Angeles
*Papers*: MARTIN C. MCGUIRE, University of Maryland
Jurisdictional Equilibrium, Group Heterogeneity, and Discriminatory Taxation
PETER J. COUGHLIN, University of Maryland, AND MELVIN J. HINICH, University of Texas-Austin
Necessary and Sufficient Conditions for Single-Peakedness in Public Economic Models
VIRGINIA WILCOX-GÖK, University of Florida
Price-Excludable Public Goods and the Core
HIROFUMI SHIBATA, Osaka University
Economics of Representative Democracy
*Discussants*: MYRNA WOODERS, University of Toronto
RADU FILIMON, Tulane University

10:15 A.M. UNEMPLOYMENT, YOUTH, AND CRIME
*Presiding*: RAY MARSHALL, University of Texas-Austin
*Papers*: W. KIP VISCUSI, Duke University
Crime as an Alternative Income Source
CHARLES MALLAR, Mathematica Policy Research, IRVING PILIAVIN, University of Wisconsin-Madison, AND CRAIG THORNTON, Mathematica Policy Research
The Perceived Benefits and Costs of Crime and their Effects on Subsequent Criminal Activity
RICHARD M. MCGAHEY, New York University
Youth Crime and Youth Employment Programs in a Changing Urban Economy
GEORGE LOEWENSTEIN, AND JAMES CATALDO, Vera Institute of Justice
Labor Market and Criminal Activities of Inner-City Youths
*Discussants*: ROBERT TAGGART, Remediation and Training Institute
PHILIP COOK, Duke University

10:15 A.M. COMPENSATION AND EMPLOYMENT DISCRIMINATION: SOME EMPIRICAL FINDINGS
*Presiding*: MARIANNE A. FERBER, University of Illinois-Urbana
*Papers*: DAVID MCCARTHY AND JOHN A. TURNER, U.S. Department of Labor
Wage Comparisons Over Time and Race/Gender Differences in Pension Benefits
TIMOTHY HANNAN, Arizona State University AND ORLEY C. ASHENFELTER, Princeton University
Sex Discrimination and Market Concentration: The Case of the Banking Industry
THOMAS F. D'AMICO, St. Peter's College
The Earnings Effects of Racial Discrimination Within and Beyond Labor Market Boundaries
*Discussants*: KENNETH J. ARROW, Stanford University
FRANCINE D. BLAU, University of Illinois-Urbana
WILLIAM A. DARITY, JR., University of North Carolina-Chapel Hill

10:15 A.M. ECONOMIC POLICY OPTIONS OF SUCCESSION LEADERS IN COMMUNIST COUNTRIES (Joint Session with the Association for Comparative Economic Studies)
*Presiding*: MARSHALL I. GOLDMAN, Wellesley College
*Papers*: ZBIGNIEW N. FALLENBUCHL, University of Windsor
Poland After Gierek
DWIGHT PERKINS, Harvard University
China After Mao
LAURA D'ANDREA TYSON, University of California-Berkeley
Yugoslavia After Tito
MARSHALL GOLDMAN, Wellesley College
The Soviet Union After Brezhnev
*Discussant*: JOHN P. HARDT, Congressional Research Service

10:15 A.M. MARXIAN ECONOMICS AND THE ECONOMICS OF KARL MARX (Joint Session with the History of Economics Society)
*Presiding*: MARTIN BRONFENBRENNER, Duke University
*Papers*: JOHN E. ELLIOT, University of Southern California
Marx on Interconnections Between Alienation and Exploitation
SAMUEL HOLLANDER, University of Toronto
Marx and Malthusianism: Marx's Secular Path of Wages
JOHN E. ROEMER, University of California-Davis
On Exploitation and Domination
*Discussants*: MARTIN BRONFENBRENNER, Duke University
MURRAY WOLFSON, Oregon State University

10:15 A.M. POLICY ISSUES ON THE ROLE AND DEVELOPMENT OF SMALL AND MINORITY BUSINESS (Joint Session with the National Economic Association)
*Presiding*: GAVIN M. CHEN, U.S. Department of Commerce
*Papers*: DAVID SWINTON AND JOHN HANDY, Clark College
The Determinants of the Rate of Growth of Black-Owned Businesses
FAITH H. ANDO, JACA Corporation, AND ROBIN C. SICKLES, University of Pennsylvania
The Determinants of Growth and Failure Rates for Minority-Owned and Non-Minority-Owned Firms
RICHARD STEVENS, U.S. Department of Commerce
Measuring Minority Business Formations and Failures
*Discussants*: LUCY REUBEN, Duke University and Federal Reserve Board
EARL HAMILTON, Federal Reserve Board
GAVIN M. CHEN, U.S. Department of Commerce

10:15 A.M. DEREGULATION AND CHANGES IN THE FINANCIAL SERVICES INDUSTRY (Joint Session with the American Finance Association)
*Presiding*: GEORGE G. KAUFMAN, Loyola University of Chicago
*Papers*: EDWARD J. KANE, Ohio State University
Regulation, Deregulation, and Banking
TIM S. CAMPBELL, University of Southern California
Implications of Deregulation for Deposit Insurance
GEORGE G. KAUFMAN, Loyola University of Chicago, LARRY R. MOTE, AND HARVEY ROSENBLUM, Federal Reserve Bank of Chicago
The Future of Commercial Banks in a Deregulated Environment
*Discussants*: DAVID S. KIDWELL, University of Tennessee
PAUL M. HORVITZ, University of Houston
ROBERT A. EISENBEIS, University of North Carolina

10:15 A.M. INFLATION, RETIREMENT PROGRAMS, AND INTERGENERATION TRANSFERS: NEW RESEARCH ON THE DISTRIBUTION OF INCOME
*Presiding*: NANCY M. GORDON, Congressional Budget Office
*Papers*: ROBERT L. CLARK AND DANIEL A. SUMNER, North Carolina State University
Inflation and the Real Incomes of the Elderly: Recent Evidence and Expectations for the Future
DONALD COX AND FREDRIC RAINES, Washington University-St. Louis
Intergenerational Income Redistribution: A Cross-Sectional Perspective
EDWARD N. WOLFF, New York University
The Distributional Effects of Social Security and Pension Wealth
*Discussants*: OLIVIA MITCHELL, Cornell University
CORDELIA REIMERS, Hunter College
EMILY ANDREWS, Social Security Administration

10:15 A.M. SOCIAL SECURITY RIFT: SHIFTS IN THE PENSION AND WELFARE COMPONENTS (Joint Session with the Association for the Study of the Grants Economy)
*Presiding*: ROBERT H. HAVEMAN, University of Wisconsin-Madison
*Papers*: ARCHIE J. NICHOLS AND MARILYN K. SPENCER, Butler University
Federal Largess with Your Trust Fund
KATHLEEN K. SCHOLL, USDA Family Economic Research Group
Farm Families: Inequities in their Social Security Program

DANIEL R. FUSFELD, University of Michigan
Redistribution to the Poor and to the Rich via Social Insurance
*Discussants*: KENNETH E. BOULDING, University of Colorado
PETER S. ELEK, Villanova University
HOWARD P. TUCKMAN, Memphis State University

12:30 P.M. LUNCHEON IN HONOR OF NOBEL LAUREATE, GEORGE J. STIGLER
*Presiding*: W. ARTHUR LEWIS, Princeton University
*Speaker*: GARY S. BECKER, University of Chicago

2:15 P.M. MICROECONOMICS OF INNOVATION AND PRODUCTIVITY GROWTH*
*Presiding*: ROGER G. NOLL, California Institute of Technology
*Papers*: JENNIFER REINGANUM, California Institute of Technology
Practical Implications of Microtheoretic Models of *R&D*
M. THERESE FLAHERTY, Harvard University
Market Share, Technology Leadership, and Competition in the International Semiconductor Industry
FRANK LICHTENBERG, Columbia University
The Effect of Federal Contract Research on Company-Financed Research
*Discussants*: GLENN LOURY, Harvard University
RICHARD LEVIN, Yale University
TOM LEE, University of California-San Diego

2:15 P.M. LIMITS TO GROWTH: RESOURCES, PRODUCTIVITY, AND SOCIAL ASPECTS (Joint Session with the American Association for the Advancement of Science and the Association of Environmental and Resource Economists)*
*Presiding*: ROGER BOLTON, Williams College
*Papers*: KENNETH BOULDING, University of Colorado
Sources of Reasonable Hope for the Future
V. KERRY SMITH, Vanderbilt University, AND JOHN KRUTILLA, Resources for the Future
Resource and Environmental Constraints to Growth
MARTIN N. BAILY, The Brookings Institution
Will Productivity Growth Recover?
*Discussants*: ROBERT PINDYCK, Massachusetts Institute of Technology
JOHN HOLDREN, University of California-Berkeley

2:15 P.M. THE DETERMINANTS OF INTEREST RATES: OLD CONTROVERSIES REOPENED*
*Presiding*: EDWARD J. KANE, Ohio State University
*Papers*: ROBERT J. SHILLER AND JOHN CAMPBELL, Yale University
Risk Premia and the Term Structure of Interest Rates
V. VANCE ROLEY, University of Washington
Unanticipated Money and Interest Rates
ALBERT K. ANDO, University of Pennsylvania
Interaction Between Fiscal and Monetary Policy and the Real Rate of Interest
*Discussants*: EDWARD J. KANE, Ohio State University
FREDERIC S. MISHKIN, University of Chicago
CARL E. WALSH, Princeton University

2:15 P.M. SEX ROLES IN ECONOMIC LIFE
*Presiding*: CORDELIA W. REIMERS, Hunter College
*Papers*: SHIRLEY P. BURGGRAF, Florida A&M University
"Womenyouthandminorities" and the Case of the Missing Productivity
PATRICIA N. PANDO, Houston Baptist University
Cloudy Days in the Sunbelt: Women and Regional Economic Conditions
MARIANNE A. FERBER AND JOE L. SPAETH, University of Illinois
Gender Inequality in the Workplace
CAROL C. FETHKE, University of Iowa
An Economic Model of Asset Division in the Dissolution of Marriage
*Discussants*: MARCIA FREEDMAN, Columbia University
FRANCINE D. BLAU, University of Illinois

2:15 P.M. Social Experimentation: A Ten-Year Retrospective
*Presiding*: Alice M. Rivlin, The Brookings Institution
*Papers*: David H. Greenberg, University of Maryland-Baltimore, and Philip K. Robins, University of Miami
New Goals for Social Experiments
Joseph R. Newhouse, The Rand Corporation
Reflections on Managing the Health Insurance Experiment
Larry L. Orr, Abt Associates
The Use of Experimental Methods to Evaluate Demonstration Projects
*Discussants*: Edward M. Gramlich, University of Michigan
Thomas Glennan, The Rand Corporation
Orley Ashenfelter, Princeton University

2:15 P.M. Closure of the Structural Deficit (Joint Session with the National Tax Association)
*Presiding*: Patric H. Hendershott, Ohio State University and National Bureau of Economic Research
*Papers*: Frank de Leeuw and Thomas M. Holloway, U.S. Department of Commerce
The Growth in the Noncyclical Deficit
Harvey Galper, The Brookings Institution
Tax Policy to Close the Structural Deficit
Richard Musgrave, University of California-Santa Cruz
Revenue Needs: A Chance for Reform
*Discussants*: Gregory Ballentine, U.S. Department of the Treasury
Michael Boskin, Stanford University and National Bureau of Economic Research
David Bradford, Princeton University and National Bureau of Economic Research

2:15 P.M. Productivity Measurement in Disequilibrium
*Presiding*: Bert G. Hickman, Stanford University
*Papers*: Ben Bernanke, Stanford University
Labor Productivity Over the Business Cycle
Charles R. Hulten, The Urban Institute and Bureau of Labor Statistics
Short-Run and Long-Run Cost Functions and the Measurement of Efficiency Change
Catherine J. Morrison, Tufts University
Productivity Measurement in Temporary Equilibrium with Nonstatic Expectations
J. R. Norsworthy, Bureau of the Census
Effects of Changing Attitudes of Workers on Production Costs: An Empirical Examination
*Discussants*: Daniel McFadden, Massachusetts Institute of Technology
Frank Brechling, University of Maryland

2:15 P.M. Economics of the Current Account
*Presiding*: Roger M. Kubarych, Federal Reserve Bank of New York
*Papers*: Jeffrey Sachs, Harvard University
Current Account Adjustments and the Developing Countries
Lars E. O. Svensson, University of Stockholm
The Terms of Trade and the Current Account
Shafiqul Islam, Federal Reserve Bank of New York, and Sweder van Wijnbergen, The World Bank
Current Account Adjustments in Major Industrial Countries
*Discussants*: Jonathan Eaton, Yale University
Jorge Braga de Macedo, Princeton University
Dale W. Henderson, Federal Reserve Board

2:15 P.M. The State of Economic Journalism
*Presiding*: Leonard Silk, *The New York Times*
*Panel*: Dean Craufurd Goodwin, Duke University
Alicia Munnell, Federal Reserve Bank of Boston
Richard Janssen, *Business Week*
David Warsh, *The Boston Globe*

2:15 P.M. Savings, Productivity and Growth: Asian Countries
*Presiding*: Lawrence R. Klein, University of Pennsylvania
*Papers*: Kazuo Sato, State University of New York-Buffalo
Savings, Investment, and Productivity Growth: A Comparative Analysis of the United States and Japan

ANDREW MASON, East-West Population Institute at Hawaii
Population Growth, Economic Growth, and Saving: A Comparison of Asian Countries and the United States
JANG H. YOO, Virginia Commonwealth University
Does Korea Trace Japan's Footsteps? A Macroeconomic Appraisal
*Discussants*: K. YAMAMURA, University of Washington
RICHARD F. KOSOBUD, University of Illinois at Chicago Circle
VINCENT SU, Baruch College, City University of New York

2:15 P.M. CLASSICAL AND NEOCLASSICAL MONETARY ECONOMICS (Joint Session with the History of Economics Society)
*Presiding*: CHARLES F. PEAKE, University of Maryland
*Papers*: ARIEH ARNON, Technion-Israel
Free Trade vs. Central Banking: The Real Bills Doctrine from Acceptance to Rejection
MICHAEL D. BORDO, University of South Carolina-Columbia
The Gold Standard: The Traditional Approach
THOMAS F. KOMPAS, Iowa State University
Walrasian and Classical Monetary Theory Reconsidered
*Discussants*: CHARLES F. PEAKE, University of Maryland
MILTON FRIEDMAN, Hoover Institution

2:15 P.M. GANDHIAN ECONOMICS: THE GRANT AND EXCHANGE INTERACTIONS (Joint Session with the Association for the Study of the Grants Economy and the Association of Indian Economic Studies)
*Presiding*: ABRAM BERGSON, Harvard University
*Papers*: ROMESH DIWAN, Rensselaer Polytechnic Institute
Gandhian Economics and the Conventional Income Distribution Models
SUSHILA J. GIDWANI, Manhattan College
Gandhian Economics and the Conventional Labor Economics
JANOS HORVATH, Butler University
Searching for the Optimal Production Scale: Some Food Supply Data
*Discussants*: PAUL JONAS, University of New Mexico
ROBERT A. SOLO, Michigan State University

2:15 P.M. INDUSTRY STRUCTURE AND STRATEGIC BEHAVIOR
*Presiding*: RICHARD SCHMALENSEE, Massachusetts Institute of Technology
*Papers*: WILLIAM P. HALL, U.S. Department of Justice
The Effects of the Learning Curve on Strategic Firm Behavior and Economic Performance
JEFFREY BALDANI, Colgate University
Advertising and Credible Entry Deterrence: A General Theory
M. ANISUL ISLAM, University of Alberta
Monopoly and Competition in the Real Estate Brokerage Industry in Canada
*Discussants*: DREW FUDENBERG, University of California-Berkeley
GARTH SALONER, Massachusetts Institute of Technology
TOMOTHY BRESNAHAN, Stanford University

2:15 P.M. JOHN R. COMMONS AWARD LECTURE (Joint Session with Omicron Delta Epsilon)
*Presiding*: ALAN A. BROWN, University of Windsor
*Paper*: DALE W. JORGENSON, Harvard University
Is Welfare Economics Impossible?

4:30 P.M. KEYNES, SCHUMPETER, MARX CENTENARY SESSION*
*Presiding*: KENNETH J. ARROW, Stanford University
*Papers*: DON PATINKIN, The Hebrew University of Jerusalem
Keynes
HERBERT GIERSCH, University of Kiel
Schumpeter
JOHN GURLEY, Stanford University
Marx

8:00 P.M. PRESIDENTIAL ADDRESS
*Presiding*: MOSES ABRAMOVITZ, Stanford University
*Speaker*: W. ARTHUR LEWIS, Princeton University

9:30 P.M. BUSINESS MEETING

**Friday, December 30, 1983**

8:00 A.M. Evidence on the Formation of Expectations
*Presiding*: Michael Darby, University of California-Los Angeles
*Papers*: Colin Camerer, Northwestern University
Rational Price Bubbles in Experimental Markets
Gerald Carlino and Robert DeFina, Federal Reserve Bank of Philadelphia
Inflationary Expectations and the Consumer
Raymond P. H. Fishe, University of Miami
Cross-Sectional Evidence on Differential Price Expectations
*Discussants*: David M. Grether, California Insititute of Technology
Stephen Figlewski, New York University
Donald J. Mullineaux, Federal Reserve Bank of Philadelphia

8:00 A.M. Current Research on Academic Labor Markets
*Presiding*: Howard P. Tuckman, Memphis State University
*Papers*: William E. Becker, Indiana University
Preselection Pools and the Quality of Economists
John Minter, John Minter and Associates
Trends in Faculty Earnings and Departmental Conditions in Economics and Related Disciplines
Howard P. Tuckman and Cyril Chang, Memphis State University
Own-Price and Cross-Elasticity of Demand for College Faculty
*Discussants*: Peter B. Meyer, Pennsylvania State University
Howard Bowen, Claremont College
Darrell R. Lewis, University of Minnesota
Charles Waldauer, Widener University

8:00 A.M. Industrial Fluctuations and Structural Shifts: Models of Schumpeter's Process of Creative Destruction
*Presiding*: Asim Erdilek, National Science Foundation
*Papers*: Gerhard O. Mensch, Case Western Reserve University
The Bi-Equilibrium Approach to Cyclical and Structural Changes in Production, Prices, and Strategic Investments
Wolfgang Weidlich, Stuttgart University
The Decomposition of Short and Long Motions: The Alternator-Coordinator Model of Industrial Fluctuation
Herman O. Wold, Uppsala University and University de Geneve
Soft Modeling Analysis of the Extended Accelerator Theory of Bias Shifts in Strategic Investment 1950–81
*Discussants*: Nguyen T. Quan, Case Western Reserve University
Werner Gerstenberger, The IFO Institute, Munich

8:00 A.M. Some Determinants of the Effectiveness of Monetary and Fiscal Policy
*Presiding*: Thomas Mayer, University of California-Davis
*Papers*: Dennis Hoffman and Don Schlagenhauf, Arizona State University
Real Interest Rates, Anticipated Inflation and Unanticipated Money: A Multi-Country Study
James Gapinski, Florida State University
Capital Malleability, Macro Performance, and Policy Effectiveness
Charles Engel, University of Virginia
Reliability of Policy Announcements and the Effects of Monetary Policy
*Discussants*: Thomas Cargill, University of Nevada
Bernard Saffran, Swarthmore College
Steven Sheffrin, University of California-Davis

8:00 A.M. Some Dynamics of Changes in the Size and Composition of Public Expenditures
*Presiding*: Robert W. Hartman, Congressional Budget Office
*Papers*: Shawna Grosskopf, Southern Illinois University, and Kathy Hayes, Northern Illinois University
The Fiscal Impact of State Mandates and Municipal Pension Reform
Randall S. Brown, Edward S. Cavin, Mathematica Policy Research, and Richard J. Murnane, Yale University
The Cost Consequences of Declining Enrollments in Public Schools
Morris Beck, Rutgers University
Theory and Evidence on Public Sector Growth
Paul N. Van de Water and Kathy A. Ruffing, Congressional Budget Office
Federal Debt and Interest Costs

*Discussants*: HELEN F. LADD, Harvard University
ERIC HANUSHEK, University of Rochester
DARWIN G. JOHNSON, Office of Management and Budget

8:00 A.M. POLICIES TO CHANGE LABOR MARKET OUTCOMES
*Presiding*: CLAIR BROWN, University of California-Berkeley
*Papers*: DEBORAH DENNO, University of Pennsylvania
Early Socioeconomic and Developmental Effects on IQ and Achievement: New Evidence
WERNER SCHINK, DAVID STERN, AND DEAN TIBBS, University of California-Berkeley
Protecting Specific Human Capital in a Recession: The California Work-Sharing Program
MAHMOOD A. ZAIDI, University of Minnesota
Incomes Policies: A Comparative Study of Australia, Canada, and the United States
*Discussants*: MYRA STROBER, Stanford University
SEAN FLAHERTY, Franklin and Marshall College
DANIEL MITCHELL, University of California-Los Angeles

8:00 A.M. ECONOMIC DEVELOPMENT, ECONOMIC EFFICIENCY: ASIAN COUNTRIES
*Presiding*: MANORANJAN DUTTA, Rutgers University
*Papers*: ANTHONY M. TANG, CLIFF J. HUANG, Vanderbilt University, AND FAQIR S. BAGI, Tennessee State University
Two Views of Efficiency in Indian Agriculture
RICHARD GRABOWSKI, DAVID SIVAN, AND RONALD L. TRACY, Southern Illinois University
Technical Change in Taiwanese Agriculture
ANINDYA DATTA, Plymouth State College-New Hampshire
Lewis Model and Development Experience: With Special Reference To India
*Discussants*: PAN A. YOTOPOULOS, Stanford University
YIEN-I TU, University of Arkansas
WALTER C. NEALE, University of Tennessee

8:00 A.M. OUTLOOK FOR FINANCIAL MARKETS (Joint Session with the National Association of Business Economists)
*Presiding*: A. NICHOLAS FILIPPELLO, Monsanto Company
*Papers*: JOHN PAULUS, Morgan Stanley, Inc.
Outlook for Interest Rates and Monetary Policy
RICHARD D. RIPPE, Dean Witter Reynolds, Inc.
Prospects for Corporate Profits
EDMOND D. VILLANI, Scudder, Stevens and Clark
Environment for the Stock Market
*Discussants*: BEN E. LADEN, T. Rowe Price Associates, Inc.
THOMAS D. THOMSON, Crocker National Bank

8:00 A.M. NEW PROSPECTIVES ON THE OPERATION OF HOUSING MARKETS. (Joint Session with the American Real Estate and Urban Economics Association)
*Presiding*: DANIEL L. MCFADDEN, Massachusetts Institute of Technology
*Papers*: KENNETH ROSEN, University of California-Berkeley, AND HARVEY ROSEN, Princeton University
Housing Tenure, Uncertainty, and Taxation
AXEL BOERSCH-SUPAN, Massachusetts Institute of Technology
Household Formation and Housing Demand
RAY STRUYK, The Urban Institute
Exploring Housing Market Behavior with a Simulation Model
*Discussants*: KONRAD STAHL, University of Dortmund
JOEL HOROWITZ, University of Iowa
DANIEL MCFADDEN, Massachusetts Institute of Technology

8:00 A.M. NEW DIRECTIONS IN MANAGERIAL ECONOMICS
*Presiding*: MARK HIRSCHEY, University of Wisconsin-Madison
*Papers*: RICHARD P. RUMELT, University of California-Los Angeles
A Model of the Entrepreneurial Firm
KEITH B. LEFFLER, University of Washington
The Effects of Entry on Product Prices
B. PETER PASHIGIAN, University of Chicago
The Effect of Environmental Regulations on Optimal Plant Size and Returns to Capital
*Discussants*: VICTOR P. GOLDBERG, Northwestern University
LYNN PARINGER, California State University-Hayward
MARTIN B. ZIMMERMAN, University of Michigan

8:00 A.M. ECONOMIC DISCONTINUITIES: POLICY RESPONSE AND INCOME TRANSFERS (Joint Session with the Association for the Study of the Grants Economy)
*Presiding*: WALTER ADAMS, Michigan State University
*Papers*: RICHARD K. BARTEL, *Challenge Magazine*
Economic Discontinuities in the 1970s: The Oil Crisis and Policy Response
RONALD MÜLLER, American University
Economic Discontinuity and Forecasting
HAROLD F. BREIMYER, University of Missouri-Columbia
U.S. Agricultural Policy and Economic Discontinuity
RICHARD S. DALE, The Brookings Institution
International Banking Innovations and Market Behavior
*Discussant*: LLOYD D. ORR, Indiana University

8:00 A.M. FUTURE DIRECTIONS FOR U.S. FOOD, AGRICULTURAL, AND TRADE POLICY (Joint Session with the American Agricultural Economics Association)
*Presiding*: WILLIAM H. MEYERS, Iowa State University
*Papers*: BRUCE BULLOCK, University of Missouri-Columbia
Future Directions for Agricultural Policy
BRUCE F. HALL, Cornell University
Future Directions for Food Policy
C. EDWARD SCHUH, University of Minnesota
Future Directions for Food and Agricultural Trade
*Discussants*: WILLIAM D. DOBSON, Purdue University
DALE HEIEN, University of California-Davis
TIMOTHY E. JOSLING, Stanford University

10:15 A.M. INSTITUTIONAL CHANGES AND THE BEHAVIOR OF INTEREST RATES
*Presiding*: BRUCE SMITH, Federal Reserve Bank of Minneapolis
*Papers*: LORETO PETER ALONZI, Loyola University of Chicago
Financial Futures, Macroequilibrium, and the Potency of Monetary Policy
HEYDAR POURIAN, University of Missouri-St. Louis
Banking Legislation and the Interest Rate Puzzles of 1930s and 1980s
ELINOR H. SOLOMON, U.S. Department of Justice
Business Loan Rate Conventions and Inflexibilities: A New Look at the Keynesian Full Employment Dilemma
*Discussants*: STEPHEN LEROY, University of California-Santa Barbara
DAVID MODEST, Columbia University
SCOTT ULMAN, University of Minnesota

10:15 A.M. LAW AND ECONOMICS: COMPENSATION OF OCCUPATIONAL DISEASE
*Presiding*: FRANCINE F. RABINOVITZ, University of Southern California
*Papers*: LESLIE I. BODEN AND CAROL A. JONES, Harvard School of Public Health
Product Liability Litigation and Workers' Compensation for Asbestos-Related Disease
PATRICIA MUNCH DANZON, Hoover Institution
The Role of Government in Liability Insurance Markets
RICHARD VICTOR, The Rand Corporation
Workers' Compensation, Experience Rating, and Workplace Safety
*Discussants*: FRANCINE F. RABINOVITZ, University of Southern California
MICHELLE J. WHITE, New York University Graduate School of Business
THOMAS BROWN, U.S. Department of Labor

10:15 A.M. EFFICIENCY AND TECHNOLOGICAL PROGRESS IN THE SOVIET UNION
*Presiding*: STEVEN ROSEFIELDE, University of North Carolina
*Papers*: KNOX LOVELL, University of North Carolina
Soviet Enterprise Efficiency: Level and Rate of Improvement
JUDITH THORNTON, University of Washington
The Soviet Response to Changing Fuel Cost
YASUSHI TODA, University of Florida
Causes of Soviet Growth Retardation: An Econometric Assessment
*Discussants*: ABRAM BERGSON, Harvard University
GREGORY GROSSMAN, University of California-Berkeley
JOHN MOORE, Hoover Institution

10:15 A.M. BENEFIT-COST ANALYSIS: THEORY AND MEASUREMENT
*Presiding*: R. CHARLES VARS, JR., Oregon State University
*Papers*: GLENN BLOMQUIST, University of Kentucky
Measuring the Benefits of Public Goods Provision Using Implicit and Contingent Markets
PAULINE M. IPPOLITO, Federal Trade Commission, AND RICHARD A. IPPOLITO, U.S. Department of Labor
Measuring the Value of Life from Consumer Reactions to New Information
FAITH H. ANDO, JACA Corporation
Idled Resources and the Benefit-Cost Calculus: Aerosol Uses of Chlorofluorocarbons
*Discussants*: W. DOUGLAS MORGAN, University of California-Santa Barbara
BRYAN C. CONLEY, Bryan Conley Forensic Consulting
HERMAN C. QUIRMBACH, The Rand Corporation

10:15 A.M. ECONOMIC POLICY RESPONSES TO BALANCE OF PAYMENTS AND DEBT CRISES (Joint Session with the Association for Comparative Economic Studies)
*Presiding*: LAWRENCE W. BRAINARD, Bankers Trust
*Papers*: EDWARD A. HEWETT, The Brookings Institution
Eastern Europe
THOMAS J. TREBAT, Bankers Trust
Latin America
(To be announced)
Asia
*Discussants*: H. ROBERT HELLER, Bank of America
ANNE O. KRUEGER, The World Bank

10:15 A.M. VALUE THEORY IN ECONOMICS (Joint Session with the Association for Evolutionary Economics)
*Presiding*: WALTER C. NEALE, University of Tennessee
*Papers*: LEWIS E. HILL AND DONALD OWEN, Texas Tech University
The Instrumental Philosophy of Economic History and the Institutionalist Theory of Normative Value
JERRY L. PETR, University of Nebraska
An Assault on the Citadel: Is a Constructive Synthesis Feasible?
ANTHONY SCAPERLANDA, Northern Illinois University
Institutionalist and Religious Values in Economic Policy Formation
JAMES I. STURGEON, University of Missouri-Kansas City
Induction and Instrumentalism in Institutional Thought
*Discussants*: WENDELL GORDON, University of Texas
MARC R. TOOL, California State University-Sacramento

10:15 A.M. DETERMINANTS OF HEALTH STATUS (Joint Session with the Health Economics Research Organization)
*Presiding*: PAUL J. TAUBMAN, University of Pennsylvania
*Papers*: HARRIET DULEEP, Social Security Administration
Socioeconomic Status and Mortality
VIRGINIA L. WILCOX-GÖK, University of Florida
A Variance Components Model of Health Determination
BARBARA WOLFE AND ROBERT HAVEMAN, University of Wisconsin-Madison
Work Behavior and Health Status
*Discussants*: MARK R. ROSENZWEIG, University of Minnesota
JOEL W. HAY, University of Connecticut Health Center
MAUREEN CROPPER, University of Maryland

10:15 A.M. PARTICIPATORY AND LABOR MANAGED FIRMS
*Presiding*: DEREK C. JONES, Hamilton College
*Papers*: LAURA D'ANDREA TYSON AND SHERMAN ROBINSON, University of California-Berkeley
The Macroeconomic Situation in Yugoslavia and the Behavior of the Self-Managed Firm
JOHN P. BONIN, Wesleyan University
Membership and Employment in an Egalitarian Cooperative
JAN SVEJNAR, Cornell University, AND DEREK JONES, Hamilton College
The Effects of Participation on Performance in Worker Cooperatives
*Discussants*: HAJIME MIYAZAKI, Stanford University
LOUIS PUTTERMAN, Brown University
STEPHEN SACKS, University of Connecticut

10:15 A.M. SIMULATING THE MACROECONOMIC EFFECTS OF ENERGY SHOCKS: A COMPARATIVE MODEL ANALYSIS BY THE STANFORD ENERGY MODELING FORUM
*Presiding*: ROBERT M. COEN, Northwestern University
*Papers*: HILLARD G. HUNTINGTON, Stanford University
A Summary of the Findings of the EMF7 Working Group on the Macroeconomic Effects of Energy Shocks
BERT G. HICKMAN, Stanford University
Key Behavioral Relationships and Parameters Affecting Macroeconomic Responses to Energy Shocks
PETER HOOPER, Federal Reserve Board, AND PETER PAULY, University of Pennsylvania
Impact of an Oil Price Shock on the U.S. Economy: International Repercussions
*Discussants*: ERNST R. BERNDT, Massachusetts Institute of Technology
MICHAEL R. DARBY, University of California-Los Angeles
JAMES D. HAMILTON, University of Virginia

10:15 A.M. CYCLICAL CHARACTERISTICS OF THE MACRO ECONOMY
*Presiding*: JAMES L. PIERCE, University of California-Berkeley
*Papers*: MICHAEL A. BERNSTEIN, Princeton University
Long-Term Growth and Cyclical Instability in the United States: Some Lessons from the 1930s
JOE PEEK, Boston College, AND JAMES A. WILCOX, University of California-Berkeley
The Effects of Taxes, Monetary Policy Rules, and Deregulation
R. G. KING, C. I. PLOSSER, University of Rochester, AND E. C. PRESCOTT, University of Minnesota
General Equilibrium and Business Cycles
*Discussants*: BEN S. BERNANKE, Stanford University
CHARLES ENGEL, University of Virginia
ROGER N. CRAINE, University of California-Berkeley

10:15 A.M. SOME MACROECONOMIC ASPECTS OF DEVELOPMENT POLICY
*Presiding*: BENT HANSEN, University of California-Berkeley
*Papers*: RICARDO MARTIN AND MARCELO SELOWSKY, The World Bank
On the Optimal Phasing of Consumer Price Subsidies
HAN X. VO, Winthrop College
A Framework for Analyzing Host Country Imports of Foreign Direct Investments
AMAR K. PARAI, Case Western University
An Analysis of Foreign Aid and Balance of Payments of the Less Developed Countries
WILLIAM DARITY, JR., University of North Carolina-Chapel Hill
Can the Poor Nations Ever Catch Up?
*Discussants*: JEFFREY NUGENT, University of Southern California
LOVELL JARVIS, University of California-Berkeley
DEBRAJ RAY, Stanford University

10:15 A.M. THE ECONOMICS OF INVENTORIES (Joint Session with the International Society for Inventory Research)
*Presiding*: KENNETH J. ARROW, Stanford University
*Papers*: RICARDO FIORITO, Università de Teramo
Inventory Fluctuations: Some Policy Experiments with a Macro Model of Italy
OWEN IRVINE, Michigan State University
Inventory Investment and Business Cycle in the United Kingdom
STEPHEN BRAUN, Federal Reserve Board
The Effects of Experience-Rated Unemployment Benefits and Inventory Costs on Production Smoothing and Labor Hoarding
*Discussants*: RICHARD E. ERICSON, Harvard University
FRANK DE LEEUW, Bureau of Economic Analysis
LOUIS MACCINI, The Johns Hopkins University

Editor's Note:
*Papers from sessions marked with an asterisk will be published in the *Papers and Proceedings* issue of the *Review*, May 1984.

# NOTES

The ninety-sixth annual meeting of the American Economic Association will be held in San Francisco, California, December 28–30, 1983.

The Professional Placement Service will be located at the St. Francis Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30.

---

Members wishing to give papers or make suggestions for the program for the AEA meeting, Dallas, Texas, December 28–30, 1984, are invited to write to Dr. Charles P. Kindleberger, Box 306, Lincoln Center, Massachusetts 01773. To be considered for the contributed sessions, abstracts of (noneconometric) papers must be received no later than February 1, 1984.

---

*Call for Papers*: The first annual meeting of the Association of Managerial Economists (AME) will be held in San Francisco, December 28–30, 1983. Two sessions of contributed papers will be featured in conjunction with the Allied Social Sciences Associations meetings. Theoretical, empirical and policy studies across a broad spectrum of topics are invited. Areas of interest include Forecasting, Pricing, Advertising, Diversification, Competitive Strategy, Innovation, Financial and Location Decisions, Regulation and Antitrust Policy. Both members and nonmembers are invited to submit papers and/or make program suggestions to the Program Chairman, Professor Mark Hirschey, Graduate School of Business, University of Wisconsin, Madison, WI 53706.

---

*Call for Papers*: The Eastern Finance Association will hold its annual meeting at Walt Disney World/EPCOT, Orlando, Florida on April 11–14, 1984. All academicians, business professionals, and government specialists with an interest in finance are cordially invited to participate. Send a two-page, single spaced abstract of your paper, indicate your interest in organizing a panel discussion, or in serving as a chairperson/discussant before October 31, 1983 to Steve Bolten, Vice President-EFA Program, College of Business Administration, University of South Florida, Tampa, FL 33620.

---

*Call for Papers*: The U.S. Army Research Institute for the Behavioral and Social Sciences will sponsor a conference on "Army Economic Manpower Policy Research," in Williamsburg, VA, December 5–7, 1984. Candidate topics for the sessions include: Determinants of Enlistment/Reenlistment; Issues in Military Compensation; Determining Army Manpower Requirements; Measuring Returns to Army Training; and Measuring the Cost Effectiveness of Weapons Systems from a Manpower Viewpoint. Research of other services is welcome if it has policy implications for the Army. The conference proceedings will be published. Send three copies of a 500-word abstract by December 1, 1983 to Dr. Curtis Gilroy or Dr. Charles Dale, U.S. Army Research Institute, 5001 Eisenhower Avenue, PERI-GP, Alexandria, VA 22333.

---

*Call for Papers*: The Southwestern Economics Association will meet on March 21–24, 1984 with the sixty-second annual meeting of the Southwestern Social Science Association. Proposals should be sent prior to November 1 to Joe C. Davis, President, Southwestern Economics Association, Department of Economics, Trinity University, 715 Stadium Drive, San Antonio, TX 78284.

---

*Call for Papers*: The 1984 International Symposium on Forecasting sponsored by the *Journal of Forecasting* and the *Manchester Business School*, will be held at the London Business School, July 8–11, 1984. Abstracts are due by February 1, 1984 and should be sent to the appropriate stream organizer on the following subjects: Econometric Forecasting; Time-Series Methods; Social and Technological Forecasting; Macroeconomic Forecasting; Judgement; Financial and Business Applications. Send abstracts and inquiries to Robert Fildes, General Chairperson, Manchester Business School, Booth Street West, Manchester M15 6PB.

---

*Call for Papers*: To commemorate the centennial anniversary of Eleanor Roosevelt's birth, a major interdisciplinary conference is planned for Vassar College on October 14–16, 1984. The purposes of the conference, cosponsored by Vassar College and the Eleanor Roosevelt Institute, are to examine the historic, social and economic event of Eleanor Roosevelt's era and her role in them, and to stimulate informed and thoughtful discussion of the applicability of Eleanor Roosevelt's vision to current domestic and international problems. Proposals describing the content, sources, methodology, and conclusions of papers should be submitted by February 1, 1984, final papers will be due August 15, 1984. Proposals and inquiries should be directed to Organizing Committee, Eleanor Roosevelt Centennial Conference, Box 186, Vassar College, Poughkeepsie, NY 12601.

---

*Call for Papers*: The third annual Presidential Conference, "Dwight D. Eisenhower: Soldier, President, Statesman," will be held March 29–31, 1984. Papers must be submitted in duplicate. Deadline is December 1, 1983. The Conference Director is Eric J. Schmertz, Dean, Hofstra University School of Law. For information, contact Natalie Datlof and Alexej Ugrinsky, University Center for Cultural and Intercultural Studies, Hofstra University, Hempstead, NY 11550. (Telephone: 516+560-5669 or 5670.)

*Call for Papers*: *Aussenwirtschaft* is a quarterly journal in international economics, addressed to national and international policymakers and management in private and public administration as well as scholars and students in international economics. Submit manuscripts, not to exceed 25 pages (double spacing, wide margins), to *Aussenwirtschaft*, Redaktionssekretar, Dufourstr. 48, CH-9000, St. Gallen, Switzerland.

*Call for Papers*: *Social Choice and Welfare*, an international journal on welfare economics, social choice, and voting theory will be published three times a year by Springer-Verlag, Heidelberg, Berlin and New York, beginning 1984. It will contain original articles, survey papers, and book reviews on the ethical and positive aspects of welfare economics and social choice. The managing editors are Wulf Gaertner, University of Osnabrück; Prasanta K. Pattanaik, University of Birmingham; and Maurice Salles, University of Caen. Submit manuscripts (four copies) to Maurice Salles, Social Choice and Welfare, Institut des Sciences de la Décision, Université de Caen, 14032 Caen Cédex, France.

*Call for Papers*: *Studies in Economic Analysis* is a biannual journal currently expanding its scope and hence soliciting research articles from both established economists and students. There is no submission fee; submit manuscripts to, or request complete format and style requirements from, The Editors, *Studies in Economic Analysis*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

Drew University is sponsoring a one-day conference on November 11, 1983 to celebrate the centenary of John Maynard Keynes. Participants are G. D. N. Worswick, Paul Davidson, Orley Ashenfelter, Leonard Silk, Robert Solow, and Lorie Tarshis. All are welcome to attend. For further information, write or call Professor Rosalind Seneca, Chair, Economics Department, HSC-52, Drew University, Madison, NJ 07940. (Telephone: 201+377-3000, Ext. 433.)

The fifth annual Research Conference of the Association for Public Policy Analysis and Management will be held in Philadelphia, October 20-22, 1983. For information about proposed panels or papers, contact Program Chairman Robert D. Behn, Institute of Policy Sciences and Public Affairs, Duke University, 4875 Duke Station, Durham, NC 27706.

The Center for Latin American and Iberian Studies at Vanderbilt University, in conjunction with the Universidad de los Andes (Bogotá, Colombia) and the Fundação Instituto de Pesquisas Econômicas of the University of São Paulo (Brazil), is holding an International Conference on Models of Political and Economic Change in Latin America in Nashville, November 3–5, 1983. The conference will address the interrelationships among political liberty, economic growth, and social equity in the context of different Latin American countries (Brazil, Chile, Colombia, Costa Rica, Mexico, Nicaragua, and Peru). For further information, contact Enrique Pupo-Walker, Director, Center for Latin American and Iberian Studies (Telephone: 615+322-2527), Samuel Morley, Department of Economics, or Jonathan Hartlyn, Department of Political Science, Vanderbilt University, Nashville, TN 37235.

*Fellowships*: The National Humanities Center will admit approximately forty Fellows for the academic year, 1984–85. Fellowships are awarded on the basis of open competition, and applications are welcomed from the United States and abroad. Representatives of the humanities, natural sciences, social sciences, and professional life may apply. The deadline is December 10, 1983. Information and application forms may be obtained from the National Humanities Center, 7 Alexander Drive, Research Triangle Park, NC 27709.

The fifteenth Latin American Congress of Sociology will be held in Managua, Nicaragua, October 3–8, 1983. The general theme is Popular Participation and Development Strategies in Latin America. For more information, contact Comité Organizador XV Congreso Latinoamericano de Sociología, Apartado Postal 167-C, Managua, Nicaragua, Centroamerica.

The sixteenth International Atlantic Economic Conference will be held October 6-9, 1983 in Philadelphia. For further information, contact John M. Virgo, Program Chairman, Atlantic Economic Conference, Southern Illinois University, Box 100, Edwardsville, IL 62026-1001.

The Graduate School of Industrial Administration, Carnegie-Mellon University, announces a postdoctoral fellowship program in political economy that provides an opportunity for researchers with a strong commitment to the use of mathematical or quantitative analysis in the study of politics and the interdependence of political and economic decision making. Fellows are to devote a twelve-month period of residence to research. There are no teaching duties. Fellows may also take advantage of GSIA's doctoral program to obtain additional training in advanced topics. Applications are invited from recent Ph.D.s in economics, political science, or related fields. All applicants should expect to receive the doctorate prior to the beginning of residency or to have received the doctorate within the past four years. Special consideration will be given to scholars interested in pursuing collaborative research with GSIA faculty in studying political processes that relate to government spending and taxation and in studying coalition formation and voting. A twelve-month stipend, fully commensurate with an assistant professor's salary, will be provided. Applicants may either forward a resume and a brief statement of research interests or write for further information to Professor Howard Rosenthal, GSIA, Carnegie-Mellon University, Pittsburgh, PA 15213.

---

The Board of Governors of the Federal Reserve System Minority Doctoral Fellowship Program provides financial support for graduate education leading to the Ph.D. to racial or ethnic minority individuals who have demonstrated high potential in economics or finance. Fellowships are awarded each spring and the recipients are selected competitively on the basis of scholastic achievement, faculty recommendations, financial need, and evidence of potential for successful completion of the doctoral work in the areas of economics or finance of concern to the Federal Reserve Board. Some preference will be given to those who have already completed one or more years of their graduate education, but first-year students are eligible. Fellowship awards typically include $450 for books, fees, and supplies; a $550 stipend for each month of the academic year and/or tuition; and the option of summer employment at the Board. For further information contact the Board of Governors of the Federal Reserve System, Minority Doctoral Fellowship Program, Board Equal Employment Opportunity Program Office, 20th & C Streets, NW, Washington, D.C. 20551. The deadline for applications is February 1, 1984.

---

Research proposals are being accepted by the Regulation and Policy Analysis Program of the National Science Foundation's Division of Social and Economics Science. The Program supports disciplinary and interdisciplinary research designed to increase and generalize knowledge about social, technical, or economic aspects of regulation. Research that examines the theoretical or actual effects of various regulatory policies on aspects of social welfare or that improves on theory, the conceptual structure of models, methodological tools, and data bases relevant to the study of regulatory phenomena is eligible for consideration, as are proposals for doctoral dissertation support, research conferences and data resources development. Dates for submission of proposals are September 1 for funding in January or February, and February 1 for funding in June or July. For further information, contact Laurence C. Rosenberg, Program Director, Regulation and Policy Analysis, NSF, 1800 G Street NW, Washington, D.C. 20550 (Telephone: 202+357-7417.)

---

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

---

## Deaths

L. Jay Atkinson, retired government economist, Washington, D.C., April 15, 1983.

R. C. Geary, former director, Economic and Social Research Institute, Dublin, Ireland, February 8, 1983.

Carl I. Van Duyne, associate professor of economics, Williams College, February 28, 1983.

## Retirement

Margaret R. Sturgis, assistant professor of economics, Auburn University-Montgomery, April 1, 1983.

Jay W. Wiley, professor, School of Management and Krannert Graduate School of Management, Purdue University, June 30, 1982.

## Promotions

A. Wayne Lacy, professor of economics, Auburn University-Montgomery, September 1, 1983.

Edward J. Ozog, assistant vice president, Open Market Operations Function, Federal Reserve Bank of New York, October 22, 1982.

Lawrence Radecki, senior economist, monetary research department, Federal Reserve Bank of New York, March 24, 1983.

### Administrative Appointments

Edward Gramlich: chairman, department of economics, University of Michigan, July 1, 1983.

John D. Scholl, University of Maryland: dean, School of Business Management, Rio Grande College, September 1982.

### Appointments

David Aschauer, University of Rochester: assistant professor of economics, University of Michigan, July 1, 1983.

Mark E. Bagnoli, Princeton University: assistant professor of economics, University of Michigan, July 1, 1983.

Michael T. Belongia, Economic Research Service, USDA: economist, Federal Reserve Bank of St. Louis, January 24, 1983.

Severin Borenstein, Massachusetts Institute of Technology: assistant professor of economics, University of Michigan, July 1, 1983.

Thomas M. Burrows, University of California-Riverside: vice president, Funds Management, Union Bank, Los Angeles, November 1982.

Jonathan A. K. Cave, University of Illinois-Urbana: associate economist, economics department, The Rand Corporation, July 1982.

Stephen Erfle, Harvard University: assistant professor of economics, University of California-Irvine, Fall 1983.

Gertrud M. Fremling, University of California-Los Angeles: visiting assistant professor of economics, University of Michigan, July 1, 1983.

Thomas H. Goodwin: assistant professor of economics, University of New Mexico, January 1, 1983.

Aaron S. Gurwitz: senior economist, regional economics staff, monetary research department, Federal Reserve Bank of New York, February 28, 1983.

Michael L. Hartzmark, University of Chicago: assistant professor of economics, University of Michigan, July 1, 1983.

Donald P. Henry, Stanford University: associate economist, economics department, The Rand Corporation, October 1982.

Gregory L. Hertzler: assistant professor of economics, Iowa State University, December 15, 1982.

David Lam, University of California-Berkeley: assistant professor of economics, University of Michigan, July 1, 1983.

Kent W. Mikkelsen, Yale University: visiting assistant professor of economics, University of Michigan, July 1, 1983.

Morris Miller: executive director, The World Bank, November 1982.

Mohan Munasinghe: senior energy advisor to the President of Sri Lanka, October 1982.

Kenneth Small, Princeton University: associate professor of economics, University of California-Irvine, Fall 1983.

V. Kerry Smith, University of North Carolina-Chapel Hill: centennial professor of economics, Vanderbilt University, September 1983.

Gary Solon, Princeton University: assistant professor of economics, University of Michigan, July 1, 1983.

Neil A. Stanley: instructor of economics, Iowa State University, January 1, 1983.

◆

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

**CAMBRIDGE**

**The Timing of Economic Activities**
**Gordon C. Winston**
This introduction to the "time specific" analysis of economic processes is firmly in the neoclassical tradition. The author applies the time-specific approach to production and uses the resulting analytical structure to examine the time-shape of efficient production of storable and perishable output, as well as the logic of duality modeling when production is explicitly seen as a process in time. $29.95

**Quantitative Economic Policies and Interactive Planning**
*A Reconstruction of the Theory of Economic Policy*
**Andrew Hughes Hallett and Hedley Rees**
By stressing the role of the policymaker rather than simply treating the economy as a physical system, Hallett and Rees focus on specifying the policy problem itself in the context of uncertainty and the variations in individual preferences, expectations, or information sets found in economic systems. $49.50

**The Theory of Natural Monopoly**
**William W. Sharkey**
An integrated presentation of the modern theory of natural monopoly that provides a useful synthesis of a subject that has been substantially transformed in recent years. Although the main emphasis is on theory, the final chapter on natural monopoly in the telecommunications industry shows the theory in use. Cloth $32.50 Paper $9.95

**The Logic of Bureaucratic Conduct**
**Albert Breton and Ronald Wintrobe**
This volume examines behavior in large organizations and the determinants of the efficiency of those organizations in government and business. Using the methods of standard neoclassical theory, the authors propose two central ideas: that members of an organization trade with one another *and* compete with one another. Their analysis breaks both with the Weberian view of bureaucracy as rational and efficient and with the view of modern public choice theory, which sees bureaucracy as inherently inefficient. $29.95

**Advances in Economic Theory** $34.50
**Advances in Econometrics** $39.50
**Werner Hildenbrand, Editor**
The papers in these collections were selected from the Fourth World Congress of the Econometric Society. They survey areas in economic theory and econometrics in which important research has appeared during the past few years.

**Limited-Dependent and Qualitative Variables in Econometrics**
**G.S. Maddala**
This book presents the econometric analysis of single equation and simultaneous equation models where the jointly dependent variables can be continuous, categorical, or truncated. Such variables are involved, for example, in models of occupational choice, choice of tenure in housing, and choice of type of schooling, as well as in models with regulated prices and rationing and models for program evaluation. $39.50

**Cambridge University Press**
32 East 57th Street • New York, N.Y. 10022

# CAMBRIDGE

## Managing Foreign Exchange Risk
*Essays Commissioned in Honor of the Centenary of the Wharton School, University of Pennsylvania*
**Richard J. Herring, Editor**

A broad and up-to-date survey of what we know about foreign exchange risk and how to cope with it. The issues addressed include the factors that give rise to foreign exchange risk, the scope for government policy to deal with it, the gains achieved by diversifying foreign exchange risk, techniques for forecasting foreign exchange rates, accounting for foreign exchange exposure, and the case for an active approach to foreign exchange management. $29.95

## Managing International Risk
*Essays Commissioned in Honor of the Centenary of the Wharton School, University of Pennsylvania*
**Richard J. Herring, Editor**

These essays provide a detailed understanding of the sources of and responses to international risk and uncertainty, bringing together the reflections of academic experts from several fields and the views of distinguished practitioners from multinational manufacturing firms, international finance institutions, and official international organizations. $29.95

## Inflation Accounting
*An Introduction to the Debate*
**Geoffrey Whittington**

A concise and clear summary of the theory of inflation accounting that describes and illustrates both theoretical and practical aspects of all the major methods.
Cloth $29.95 Paper $9.95

## Analysis Without Measurement
**Donald W. Katzner**

Katzner addresses the question of how the many phenomena in the social sciences that appear, at least in part, impossible to measure might be understood and investigated in the absence of numerical gauges. In offering a coherent approach to the analysis of the non-quantifiable, he develops static and dynamic modeling techniques for non-scaleable variables and describes their varied applications, along with methods for testing hypotheses obtained through these and other procedures.
$37.50

## The Economics of Price Discrimination
*Four Essays in Applied Price Theory*
**Louis Phlips**

This book offers a theoretical and unified explanation of how prices are determined in practice and represents the first unified treatment of a recent and growing body of literature in the "new industrial organization" field. It fills the gap between textbook microeconomics and real-life antitrust cases in a straightforward, nontechnical way. Cloth $39.50 Paper $12.95

## Intertemporal Planning, Exchange, and Macroeconomics

**D.F.I. Folkerts-Landau**

This original and rigorous effort to incorporate into economic theory transactions that impose costs upon participants provides an analysis of modern exchange and macroeconomics and general equilibrium theory. The author investigates the way individual agents trade and accumulate money, bonds, and goods over their life cycles, and his explicit introduction of trading costs in the form of a transaction technology sets his analysis apart from the available literature. $32.50

## Economic Behavior Within Organizations

**Stephen A. Hoenack**

In a comprehensive economic analysis of the internal workings of organizations, Stephen Hoenack argues that the costs of obtaining information lead managers to delegate control of resource allocation to their subordinates and that these information costs determine the effectiveness of the constraints that can be put upon subordinates' discretion. $34.50

## Econometric Society Monographs in Pure Theory

**F.H. Hahn, Editor**

**Hugo Sonnenschein, Editorial Adviser**

This new series will present up-to-date research in mathematical economics and pure theory.

## Mathematical Economics

*Twenty Papers of Gerard Debreu*

**Gerard Debreu**

Debreu has selected twenty of his papers for inclusion in this volume. They are preceded by a major introduction by Werner Hildenbrand, who assesses Professor Debreu's contribution to economic theory and explains the part played by these papers in the development of the subject. $34.50

## Money and Value

*A Reconsideration of Classical and Neoclassical Monetary Economics*

**Jean-Michel Grandmont**

This study examines one of the major theoretical issues underlying a number of recurrent controversies in macroeconomics: whether a competitive monetary economy has built-in mechanisms strong enough to remove excess supply and demand in all markets through an automatic adjustment of the price system.
About $29.95

## Disequilibrium Foundations of Equilibrium Economics

**Franklin M. Fisher**

Professor Fisher provides a review of the older literature on the stability of general equilibrium and a consideration of a more satisfactory general model in which agents realize that they are in disequilibrium and act on arbitrage opportunities.
About $34.50

**Cambridge University Press**
32 East 57th Street • New York, N.Y. 10022

# New
# from The University Press of America

**How Economists Explain: A Reader in Methodology**
*edited by* William L. Marr *and* Baldev Raj

An outstanding collection of articles on economic methodology—the concepts, theories, and basic principles of reasoning which are part and parcel of the discipline of economics. The nine parts of this text include articles by prominent economists such as F. Machlup, R. Heilbroner, A. Coddington, L. Boland, H. Simon, and W. Leontief. Includes *The Nature of Economics; Economics and Value Judgement; The Goal of Economics as a Hard Science; Should Prediction be the Goal of Economics?; The Role of Empiricism in Economics; How Theories are Verified; Some Aspects of Model Building and Policy Making Role; Pattern Modeling, Story-telling, Holism, and Behavioral Requirements; What is Right and Wrong With Economics?* Also includes a glossary of philosophical and methodological terms, author index, and bibliography.

1983, 414 pages, LC: 83-5863
Paper, 3192-X $15.75
Cloth, 3191-1 $28.00

**Economic Anthropology: Topics and Theories**
*Monographs in Economic Anthropology No. 1*
*edited by* Sutti Ortiz

Co-published with the Society for Economic Anthropology, this collection of papers was developed from the inaugural conference of the Society, with the participation of anthropologists, ecologists and political scientists. The uniformly fine papers reflect the interdisciplinary dialogue and the joint interest in a critical review of the various approaches used to examine the organization and transformation of economic systems.

1983, 446 pages, LC: 83-10296
Paper, 3322-1 $15.75
Cloth, 3321-3 $27.50

**From Basic Economics to Supply-Side Economics**
*by* Charles T. Stewart, Jr. *and* Melvin L. Greenhut

This excellent book includes formal, yet simplified, explanations of supply and demand analysis, Keynesian economics, monetarism, and in particular, supply side economics, including the current Reagan variation of it. It provides detailed explanations of our money and banking system, as well as evaluations of the economic progress and problems of other countries. The clear explanations and high readability give this volume wide spread appeal.

1983, 200 pages
Paper, 3427-9 $ 9.75
Cloth, 3426-0 $19.75

**The Inflation Crisis, and How to Resolve It,** *Second Edition*
*by* Henry Hazlitt

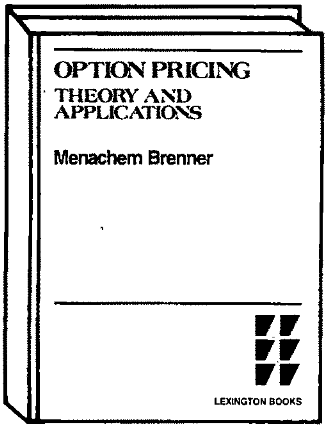
An excellent volume by an eminent journalist and author which presents a history of inflation, an explanation of its causes, an analysis of the misconceptions and fallacies that prevail about it, the outlook for more of it, and advice regarding what the reader can and cannot do to protect himself against it. The author recognizes the strength of the political forces that allow its continuance, and the urgent need of the re-establishment of an international gold standard. A first-rate analysis of an ongoing problem.

1983, 192 pages
Paper, 3301-9 $ 9.00

To order, write
**The University Press of America**
4720 Boston Way • Lanham, Maryland 20706 • 301-459-3366

*Call or write for our new Fall Economics Catalog.*

## The Latest Research in Economics



### Option Pricing

Theory and Applications
edited by **Menachem Brenner,** Hebrew University and New York University
A Salomon Brothers Center Book
Distinguished academics and professionals analyze the main theoretical and empirical developments in option pricing, the major new field of investigation in financial economics.
256pp. ISBN 0-669-05714-2 $25.95

### Fiscal Theory

Government, Inflation, and Growth
**Hans Brems,** University of Illinois
Explores the analytical foundations of Keynesian and monetarist fiscal theories and formulates a new balanced synthesis.
208pp. ISBN 0-669-05688-x $22.95

### Entrepreneurship, Productivity, and the Freedom of Information Act

Protecting Circumstantially Relevant Business Information
**William L. Casey, Jr., John E. Marthinsen,** and **Laurence S. Moss,** Babson College
Identifies circumstantially relevant business information that is both proprietary and vital to the entrepreneurial process and analyzes the protection afforded it by the Freedom of Information Act trade-secrecy exemption.
240pp. ISBN 0-669-06349-5 $24.95

### Strategies for Vertical Integration

**Kathryn Rudie Harrigan,**
Columbia University
Foreword by **William H. Newman**
*"The most comprehensive study of the important topic of vertical integration that has been undertaken in recent years. . . . The book will become a basic guide for both scholars and practitioners."* —from the foreword
400pp. ISBN 0-669-06694-x $29.95

### The Future of Nonfuel Minerals in the U.S. and World Economy

Input-Output Projections, 1980-2030
**Wassily Leontief, James C.M. Koo, Sylvia Nasar,** and **Ira Wassily Sohn,** Institute for Economic Analysis, New York University
Includes a discussion of the role of resources in the economy, a review of past U.S. mineral policy, and a summary of previous studies projecting future minerals requirements.
512pp. ISBN 0-669-06377-0 $39.95

### The Role of the State in Property Taxation

edited by **H. Clyde Reeves,** with **Scott Ellsworth**
A Lincoln Institute of Land Policy Book
Authors from sixteen states and one Canadian province describe the range of state influence on taxation and suggest using new technologies for better collection of state revenues.
240pp. ISBN 0-669-06292-8 $25.95

### The Economics of Cable Television

**G. Kent Webb**
Discusses the many telecommunications services available from cable and explores the problems of efficient pricing and system design.
224pp. ISBN 0-669-06323-1 $22.95

## from LexingtonBooks

HEATH
**LexingtonBooks,** D.C. Heath and Company
125 Spring Street, Lexington, MA 02173
(617) 862-6650 (212) 924-6460
**Call our toll-free numbers,**
(800) 428-8071 or (800) 33-HEATH

# Harper & Row presents: essentials

*NEW*

## Lipsey & Steiner

## ECONOMICS ***Seventh Edition***

Combining solid theory with real-world applications, this classic best seller covers macroeconomics and microeconomics in separate, self-contained sections. For this edition, the chapters on macroeconomic theory and applications have been extensively revamped. A full commitment to the aggregate demand-aggregate supply approach features a superlative exposition of the "supply side" of the economy. Additional coverage of rational expectations enhances Chapter 38: *Current Issues in Macroeconomic Policy.* Changes in the micro section include updated coverage of OPEC and two chapters instead of one devoted to household consumption behavior. 1/84. 1024 pages/tent. Instructor's Manual. Study Guide. Test Bank with 3000 multiple-choice questions.

*NEW*

## Hamermesh & Rees

## THE ECONOMICS OF WORK AND PAY

***Third Edition***

Virtually a new text, this significantly expanded and updated edition includes key policy issues to liven up exposition...shows the process by which labor market outcomes are produced...integrates data and empirical work directly with theory...avoids unnecessary mathematics...illustrates points with international comparisons...and shows what labor market facts are and how to explain them. Now includes new material on policy issues in a wide variety of areas (nearly 50 in all), and revamped coverage of search theory and results; tied migration; implicit labor market contracts: hedonic wages; and much more. 2/84. 550 pages/tent.

*NEW*

## Sherman & Evans

## MACROECONOMICS

**An Introduction to Keynesian, Monetarist, and Marxist Views**

This new introduction explores a variety of different economic systems—from the Marxist and institutionalist to the monetarist and supply side. It focuses on dynamic disequilibrium in a working economy and employs an empirically-based approach to give students a realistic perspective on such contemporary issues as large-scale unemployment and inflation. Features extensive new material on money and business cycles, including new research by the authors in those areas, and incorporates historical and institutionalist concepts into a broad political-economic model of the macroeconomic process. 12/83. 576 pages/tent. Instructor's Manual.

# for your economics courses.

## Waud
## ECONOMICS
***Second Edition***

Updated, expanded, and refined, this highly teachable introduction to macro and microeconomics offers such superb features as a revised and updated version of the first edition's pioneering discussion of aggregate supply and its significance for today's economic problems; new chapters on comparative economic systems, government regulation, and the economics of energy; learning aids like News Items (many adapted from the *Wall Street Journal*) and Checkpoints (with answers now at the end of the book); and a revamped Test Bank by Michael Salemi, nationally recognized authority on teaching methods in economics (all questions have been subjected to statistical test item analysis at the University of North Carolina, Chapel Hill). 1/83. 866 pages + Glossary & Index. Instructor's Manual. Transparency Masters. Study Guide. Test Bank.

## Miller
## ECONOMICS TODAY
***Fourth Edition***

This national bestseller is distinguished by its issues-and-applications orientation and a vivid writing style. Now updated to include supply-side economics and other current topics. 1982. 830 pages. The fourth edition is supported by a Total Learning/Teaching Package that insures clear teaching and understanding of the material. The Package includes a new Student Learning Manual, a new Instructor's Manual, a new class-tested Test Bank of 2000 items (available on ACCESS), 100 transparency masters, 30 audio cassettes (15 on macro, 15 on micro) with self-testing student workbooks, and a new set of 30 videotapes covering the macro view of economics, with accompanying Student Manual and Instructor's Manual.

**To request examination copies,** write to Suite 3D, Harper & Row, 10 East 53d Street, New York, N.Y. 10022. Please include course title, enrollment, and present text.

**Harper & Row**

*The rare book collection of the*

**ROYAL STATISTICAL SOCIETY**

*in microfiche*

The rare book collection of the **Royal Statistical Society** contains irreplaceable monographs which have long been needed by scholars.

Recently the Council of the Society extended an agreement to Microforms International Marketing Corporation, a subsidiary of Pergamon Press, for filming this collection and making microfiche reproductions available.

A first segment of this collection has already been filmed and is available in negative polarity, NMA standard 98-frame microfiche.

Hereunder, a sampling of abbreviated titles:

**BELL, William,** Causes which render a nation populous. Cambridge, 1756—1 fiche.
**BLACK, William,** Mortality of the human species. London, 1788—5 fiche.
**CONDORCET, M.J.A.C.,** Essai sur . . . la probabilite des decisions. Paris, 1785—6 fiche.
**DE MOIVRE, Abraham,** The doctrine of chances. London, 1718—3 fiche.
**DOUBLEDAY, Thomas,** The true law of population. London, 1853—5 fiche.
**ENSOR, George,** . . . The population of nations. London, 1818—6 fiche.
**GOMPERTZ, Benjamin,*** . . . The law of human mortality. London, 1825—1 fiche.
**GRAUNT, John,*** Natural and political observations. London, 1664—2 fiche.
**HAZLITT, William,** A reply to the *Essay on Population* of Malthus. London, 1807—5 fiche.
**HOWLETT, John,** An examination of Dr. Price's essay on population. London, 1781—2 fiche.
**JARROLD, Thomas,** Dissertation . . . in answer to the Malthus *Essay.* London, 1806—4 fiche.
**MEAD, Richard,** A discourse on the plague. London, 1744—3 fiche.
**PETTY, William,*** Five essays in political arithmetick. London, 1687—2 fiche.
**QUETELET, Lambert A. J.,*** Recherches sur la population. Brussels, 1827—2 fiche.
**SHORT, Thomas,*** . . . The increase and decrease of mankind in England and several countries abroad. London, 1767—3 fiche.
**STORCH, Heinrich F.,** Statistische Uebersicht . . . des russischen Reichs. Berlin, 1795—2 fiche.
**SUSSMILCH, Johann P.,*** Versuch . . . die erste Sprache . . . nicht vom Menschen. Berlin, 1766—2 fiche.

*Among most of their works.

Fiche are sold at US$6.00 each with a required minimum purchase of ten fiche. These prices are applicable for all countries except Japan.

For expanded listings and orders, please write to:
**Microforms International Marketing Corporation**
A Subsidiary of **Pergamon Press**
Fairview Park, Elmsford, New York 10523

6/83/G/216

## BUSINESS CYCLES, INFLATION, AND FORECASTING

### Second Edition

**Geoffrey H. Moore**

". . . an exhaustive and informative investigation into the nature of business cycles."
— *Southern Economic Journal*

The new edition of this highly praised work includes up-to-date information on recent business cycles, a signaling system to identify the beginning and ending of recessions, new findings on inflation, an update on leading indicators for the "big seven" countries, a new leading index of employment and unemployment, and new essays on current problems with economic statistics.

**July 488 pages $42.50† cloth $19.95† paperback**

## EXCHANGE RATE AND TRADE INSTABILITY

### Causes, Consequences, and Remedies

**David Bigman** and **Teizo Taya,** Editors

Economists from throughout the world investigate the variability of leading industrial nations' exchange and interest rates during the 1973-1981 period of generalized floating, contrast theories of exchange rate determination, consider conditions under which private speculation can strengthen a shaky foreign exchange market, analyze the stability of a wide range of applied monetary models, and explore the causes and effects of recent export instability.

**March 376 pages $39.95**

## STEEL: Upheaval in a Basic Industry

**Donald F. Barnett** and **Louis Schorsch**

Barnett and Schorsch scrutinize U.S. productivity levels, draw enlightening comparisons with the Japanese and Canadian steel industries, and offer new perspectives in their analysis of economic policy. Concluding that a revitalized steel industry is not an unrealistic goal, they outline its likely structure and suggest changes in corporate strategy and governmental policy necessary to improve the economic performance of this vital industry.

**September ca 336 pages $28.00†**

## INTERNATIONAL AND REGIONAL CONFLICT

### Analytic Approaches

**Walter Isard** and **Yoshimi Nagao,** Editors *(The Peace Science Studies Series)*

Using mathematical models and value-based decision rules, scholars from around the world address the field of conflict analysis. They offer new approaches to problems that range from the theoretical to the actual, the international to the personal. Rigorously scientific throughout, these essays offer concrete proposals for the resolution of conflicts at all levels.

**1983 256 pages $36.00†**

## STALEMATE IN TECHNOLOGY

### Innovations Overcome the Depression

**Gerhard O. Mensch**

"In many ways the most important contribution to macroeconomic theory since John Maynard Keynes."
— *World Press Review*

"The one person who appears to me to have reached an optimal understanding of technological innovation and its connections to economic problems is Gerhard Mensch. It is Mensch's thesis that the best way to overcome economic depression, inflation, and stagflation is through basic or radical innovations, most likely technological but perhaps social and institutional as well." — *Challenge*

**1979 272 pages $19.50 cloth $11.95 paperback**

† denotes tentative price. Prices subject to change without notice.

Ballinger PUBLISHING COMPANY *A Subsidiary of Harper & Row, Publishers, Inc.*
54 Church Street • Cambridge, MA 02138

# St. Martin's Press Fall 1983:

## South-South Financial Cooperation

edited by Dragoslav Avramovic,
*Office of the Secretary General, UNCTAD*

In an attempt to present alternatives to the current worldwide financial crisis, the highly-skilled contributors to this volume suggest the best methods for achieving a South-South cooperation in the field of finance. As a whole, the collection of essays assert that developing countries should try to utilize their own resources and establish new institutions and modalities for development and other financing.

*September 74741-1 250 pp. $27.50*

## The New International Economic Order

### Conflict and Cooperation in North-South Economic Relations, 1974–1977

Jeffrey A. Hart, *Indiana University*

In this penetrating new study of the politics of economics, Hart raises important questions concerning the series of negotiations between rich and poor nations to establish new rules for government economic relations that occurred between 1974 and 1977.

*September 56819-3 215 pp. $22.50*

## The Nigerian Oil Company

J. K. Onoh, *Imo State University, Nigeria*

Nigeria is rated as the world's tenth largest producer of oil, yet little information is available about the implications of oil production on the Nigerian economy. For the first time, the development of Nigeria's oil industry is examined in this comprehensive volume. Onoh charts the changing course of Nigeria's economy and examines the dramatic effect oil has had on her domestic and international policies.

*September 57274-3 250 pp. $27.50*

St. Martin's ISBN prefix: 0-312.

*Prices and publication dates subject to change without notice.*

## Federalism in Nigeria

S. Egite Oyovbaire,
*Ahmadu Bello University, Nigeria*

This timely publication, coinciding with the current debate in Nigeria about the organization and management of federalism, is a critical analysis of the development of the Nigerian State and the ways in which federalism has been used to further national integration. The author provides a powerful critique against the current theory of comparative federalism. Using Nigeria as a case study, the author demonstrates that federalism which exists under conditions of underdevelopment, communal pluralism and rapid social transformation is not a situation of coordinate or equal power status of the two levels of government, but a case of dependence and autonomy of the state vis-à-vis the central government in the acquisition and use of power.

*September 28551-5 340 pp. $35.00*

## Slow Growth and the Service Economy

Pascal Petit

Presenting a clear and comprehensive macroeconomic framework which accounts for supply-and-demand relations in different sectors of the economy, the author analyzes various determinants of the demand for services and the conditions which allow for a perspective on the volume and quality of employment in the service industries using data accumulated over the last thirty years. In the course of this vital argument, the role of services in economic growth is thoroughly assessed, as are the effects of the current slowdown on market forces and on economic policies.

*September 72947-2 200 pp. $25.00*

Send for our new Fall 1983 catalogue of reference and scholarly books. To order these books at 20% discount off the list price, please send your check or institutional purchase order directly to:

# Economic studies

## The West German Economy

Eric Owen Smith,
*Loughborough University of Technology*

In two centuries (1780–1980) the German nation has developed from one large geographical area of 1,789 territories into two major economic powers. To understand the collapses, turbulence and spectacular recoveries of the German economy, the author explores six crucial recurrent features: hyper-inflation and currency reform; a high rate of technical advance with a high export potential; a legal framework which was conducive to a high degree of business concentration and thus the derivation of large economies of scale; a banking sector geared to active participation in the industrial sector; and active involvement by the government in economic and social affairs.

*September* 86290-3 *350 pp.* $32.50

## The Economics of Peacemaking

Focus on the Egyptian-Israeli Situation

Ruth Arad, Seev Hirsch, *both of Tel Aviv University,* and Alfred Tovias, *Hebrew University*

Starting with the two key concepts of dependence and interdependence, this book offers an economic interpretation of the dynamics of peacemaking. Interdependence is viewed as a state characterized by an acceptable balance between gains from economic transactions between recent or potential belligerents and losses from disruption caused by dissociation between them.

*October* 23441-4 *260 pp.* $25.00

## Oil and the British Economy

Sir Fred Atkinson, *formerly chief economic advisor of Her Majesty's Treasury,* and Stephen G. Hall, *National Institute of Economic and Social Research, London*

The 1976 discovery of oil in the North Sea and the rest of the continental shelf was widely expected to usher in a period of prosperity for the British economy. Yet the early 1980s have been years of recession and unemployment. To discover the reasons behind this predicament, the authors examine the North Sea oil sector itself and show the direct impact of its recent development on the British economy. They analyze the automatic responses of an economy to the development of such a new sector, the "Kay debate," and the experience of six other countries with oil and the manufacturing industry.

*October* 58295-1 *200 pp.* $32.50

## Regional Integration in East and West

edited by Christopher T. Saunders,
*University of Sussex*

The strategy of economic integration, by which groups of nations band together for the purposes of economic cooperation is examined by leading economists from both the east and west in this stimulating study. In particular, it focuses on the Council for Mutual Economic Assistance (CMEA) of the centrally planning countries of Eastern Europe, with the Soviet Union as their joint bond-holder, and the European Economic Community (EEC) of the free market economy countries. Each of these organizations is studied in detail, including aspects of their relations with non-members, their interactions with other multinational groups, specific components of their integrational policies, and monetary issues.

*October* 66917-8 *280 pp.* $32.50

# St. Martin's Press

*Reference & Scholarly Books Division, Rm. 132*
175 Fifth Avenue, New York, New York 10010

**Basic Books, Inc.**

**Decision Making at the Top**
The Shaping of Strategic Direction
Gordon Donaldson and Jay W. Lorsch
$18.50 *September*

**Economic Development**
Ian M.D. Little
A Twentieth Century Fund Book
$22.95

**Game Theory**
A Nontechnical Introduction
Revised Edition
Morton D. Davis
$20.00 cloth, $7.95 paper *September*

**Industrial Renaissance**
Producing a Competitive Future for America
William J. Abernathy, Kim B. Clark, and Alan M. Kantrow
$19.00

**The International Money Game**
Fourth Edition, Revised and Expanded
Robert Z. Aliber
$16.00 cloth, $8.95 paper

**The Making of Modern Advertising**
Daniel Pope
$18.95

**The Open Economy and the World Economy**
John Williamson
$23.95

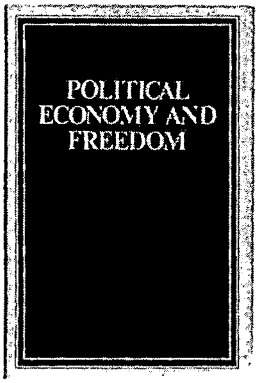
**Principles of Money, Banking, and Financial Markets**
Fourth, Revised Edition
Lawrence S. Ritter and William L. Silber
$25.95

**Structuralist Macroeconomics**
Lance Taylor
$18.95

**The Truth About Supply-Side Economics**
Michael K. Evans
$17.95

**Basic Books, Inc.**
10 East 53rd St. New York, N.Y. 10022

LibertyPress
LibertyClassics



Political Economy
and Freedom:
*A Collection of Essays*

By G. Warren Nutter
Edited by Jane Couch Nutter
Foreword by Paul Craig Roberts

G. Warren Nutter (1923-1979) was a fine economist who was also a keen student of political institutions. He was a political economist in the classical sense.

The thirty-three essays in this collection illustrate the broad range of Nutter's thought. There are essays on the Soviet economy and international relations as well as essays exploring the economic institutions which support a free society.

One finds in these essays a man of intellect and judgment ever ready to look at the evidence. He defends capitalism, not because it is perfect, but because for this imperfect world it is superior to the attainable alternatives.

Hardcover $10.00; Paperback $5.50

Prepayment is required on all orders not for resale. We pay book rate postage on prepaid orders. Please allow 4 to 6 weeks for delivery. *All* orders from outside the United States *must* be prepaid *in U.S. dollars*. To order, or for a copy of our catalogue, write:

Liberty*Press*/Liberty*Classics*
7440 North Shadeland, Dept. F101
Indianapolis, IN 46250

# Institute for International Economics

The Institute for International Economics was created in 1981 in order to provide fresh analyses of major issues in international economic policy and to suggest new approaches to dealing with the problems involved. Books and monographs published by the Institute are distributed by The MIT Press.

## International Debt

Systemic Risk and Policy Response

*William R. Cline*

This book traces the origins of international debt and recent trends that burden it. It examines the extent of the financial system's vulnerability, and the adequacy of bank regulation and of central bank coverage for emergency lending to international banks, and reviews the prospects for orderly servicing of the debt in the next few years.

175 pp. $20.00

## Trade Policy in the 1980s

*edited by William R. Cline*

The world trading system which has been of critical importance to both economic prosperity and political harmony throughout the postwar period now faces strains not seen since the 1930s. Essays in this book assess the trends and problems in trade policy, recommended ways of dealing with them.

714 pp. $35.00

## Subsidies in International Trade

*Gary Clyde Hufbauer and*

*Joanna Shelton-Erb*

Subsidies raise some of the most important and difficult issues of contemporary international economic relations. Proposals for improving the ability of the international trading system to tackle the problems subsidies pose are set forth in this book.

400 pp. $20.00

## IMF Conditionality

*edited by John Williamson*

These contributions (including eleven country studies) assess the controversy surrounding the International Monetary Fund and provide judgments about the criteria for Fund lending which should help readers understand and analyze both its ongoing role in smoothing adjustment to international payments imbalances and its currently critical position in responding to the debt crisis.

679 pp. $30.00

## Policy Analyses in International Economics

paperback monographs

**The Exchange Rate System**

*John Williamson*

80 pp. $6.00

**Economic Sanctions in Support of Foreign Policy Goals**

*Gary Clyde Hufbauer*

80 pp. $6.00

**International Debt and the Stability of the World Economy**

*William R. Cline*

80 pp. $6.00

**The Multiple Reserve Currency System**

*C. Fred Bergsten and John Williamson*

80 pp. $6.00

*Write for our current economics catalog*

28 Carleton Street, Cambridge, MA 02142

**Trade Policy in the 1980s**

*C. Fred Bergsten and William R. Cline*

84 pp. $6.00

**"Reciprocity"**

A New Approach to World Trade Policy?

*William R. Cline*

42 pp. $6.00

**The Lending Policies of the International Monetary Fund**

*John Williamson*

72 pp. $6.00

Available from Books By Phone, toll free
1-800-255-2665
(in Connecticut 1-203-966-5470)
MasterCard and Visa accepted

**THE MIT PRESS**

# New Fall Books From MIT

**Inflation, Debt, and Indexation**
*edited by Rudiger Dornbusch and*
*Mario Henrique Simonsen*

These essays cover frontier topics in modern macroeconomics and macropolicy, ranging from the highly theoretical to practical experiences in Brazil and in Europe.
360 pp. $25.00

**Global Econometrics**
Essays in Honor of Lawrence R. Klein
*edited by F. Gerard Adams and*
*Bert G. Hickman*

Contributions by Paul Samuelson, Albert Ando, Nissan Liviatan, Wilhelm Krelle, Phoebus Dhrymes, Mitsuo Saito, T. W. Anderson, R. J. Ball, and others pay tribute to the recipient of the 1980 Nobel Prize in economics. They cover economic theory, econometric methodology, applied micro and macroeconomics.
448 pp. $45.00

**Marxism, Central Planning, and the Soviet Economy**
Economic Essays in Honor
of Alexander Erlich
*edited by Padma Desai*

"This is a rich, well-edited collection of articles on centrally planned economies ranging broadly from the intricacies of Marxist theory and resource allocation analysis to the index-number theoretics of Soviet defense spending."
–Herbert S. Levine, University of Pennsylvania
312 pp. $35.00

**The Multinational Corporation in the 1980s**
*edited by Charles P. Kindleberger*
*and David B. Audretsch*

These essays address the vital question of how much of the theory of direct foreign investment, developed a decade ago, still holds in the light of the shift from fixed to flexible exchange rates, the rise of OPEC, the increasing nationalization of firms, and other changes.
384 pp. $30.00

*Original in Paperback*

---

**Common Crisis**
North-South
Cooperation for World Recovery
*The Brandt Commission*

*Common Crisis* is a call to world government for emergency measures to halt international economic collapse. In lucid and forceful terms it describes the crisis in trade, energy, and food, offering specific proposals for reversing these negative trends.
184 pp. $4.95

**Macroeconomics After Keynes**
A Reconsideration of the *General Theory*
*Victoria Chick*

This reassessment results from the author's experience in using Keynes's book as the core of her macroeconomics courses for undergraduates at University College in London because of the rich and unique understanding it gives of the structure of macroeconomic interactions and methods of analyzing them. "Ms. Chick succeeds in clearing up yet a few more errors in the conventional interpretation of Keynes, some of them quite fundamental and basically simple."
–Basil Moore, Wesleyan University
356 pp. $13.95 (cloth $30.00)

---

*Available from Books by Phone, toll free*
*1-800-255-2665*
*(in Connecticut 1-203-966-5470)*
*MasterCard and Visa accepted*

*Write for our current economics catalogue*

28 Carleton Street
Cambridge, MA 02142

**THE MIT PRESS**

# Government &Technical Progress

## A Cross-Industry Analysis

edited by *Richard R. Nelson*,
Yale University

First-of-its-kind—a study of Federal policies that have influenced—for better or worse—technical growth in seven US industries: semiconductor, computer, aircraft, pharmaceutical, construction, agriculture, and automobile. Through careful analysis and research, contributors suggest the most productive balance between contemporary technological policy and economic growth.

**Contents:** Introduction, *Richard Nelson* • The Semiconductor Industry, *Richard Levin* • The Computer Industry, *Barbara Katz* and *Almarin Phillips* • The Aircraft Industry, *David Mowery* and *Nathan Rosenberg* • The Pharmaceutical Industry, *Henry Grabowski* and *John Vernon* • Agriculture, *Robert Evenson* • Residential Construction, *John Quigley* • The Automobile Industry, *Lawrence White* • Government Stimulus of Technical Progress: Lessons from American Industry, *Richard Nelson* • Indexes

512 pp. 1982 $45.00 hardcover

A volume in *The Technology Policy and Economic Growth Series*

# Prices,Wages &Business Cycles

## A Dynamic Theory

by *Burton Klein*, California Institute of Technology

In a controversial rethinking of macroeconomics, the author presents a dynamic concept of competition as it relates to the overall performance of an industry, and makes a provocative case against protectionism and "cooperative R & D."

**Contents:** Preface • Introduction • Preview • The relationship between risk-taking at the micro level & the rate of progress at the macro level • Utilizing technological building blocks to generate productivity gains • Implications for macro models & equilibrium theory • Dynamic competition • Dynamic efficiency as a process • A model of dynamic processes • The dynamic response to foreign competition • Deadweight drag & business cycles • Determinants of changes in wage rates • The second law of dynamic economics • Conversations with Adam Smith • Indexes

250 pp. 1983 $29.50 hardcover

---

## Other titles in *The Technology Policy and Economic Growth Series*

*Herbert I. Fusfeld* and *Richard R. Nelson*, Editors

### UNDERSTANDING R & D PRODUCTIVITY

edited by *Herbert I. Fusfeld* & *Richard N. Langlois*, Center for Science and Technology Policy, New York University

168 pp. 1982 $19.50 hardcover

### THE U.S. MICROELECTRONICS INDUSTRY

Technical Change, Industry Growth, and Social Impact

by *Nico Hazewindus*, N.V. Philips, The Netherlands, with *John Tooker*

232 pp. 1982 $25.00 hardcover

### INDUSTRIAL PRODUCTIVITY AND INTERNATIONAL TECHNICAL COOPERATION

edited by *Herbert I. Fusfeld* & *Carmela S. Haklisch*, Center for Science and Technology Policy, New York University

176 pp. 1982 $19.50 hardcover

**Pergamon Press, Inc.**
Maxwell House, Fairview Park
Elmsford, New York 10523

5/83/9H/108

Praise from John Galbraith,
Irving Kristol, the *New York Times*...

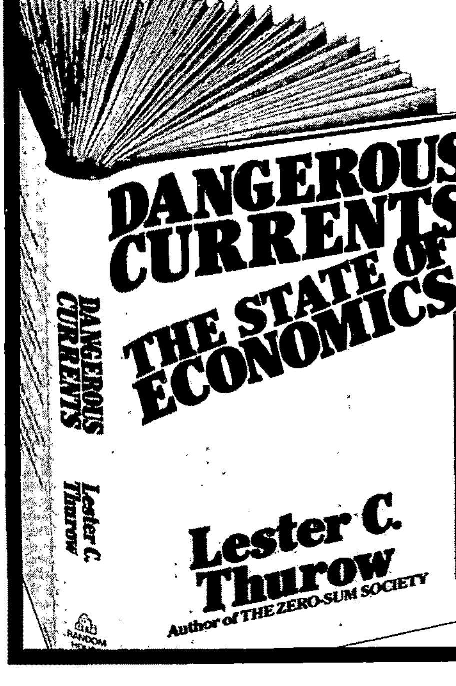
**LESTER THUROW**

**supplies "the best analysis of the source of economic problems that I have ever read."***

"**Richly informed and informative,** skillfully argued, and otherwise extremely good...Thurow has made clear–not alone, one trusts, for liberals –what was so wrong."
–JOHN KENNETH GALBRAITH, *The Atlantic*

"**What a splendid critique it is! Bold, incisive, original.**" –IRVING KRISTOL, *New Republic*

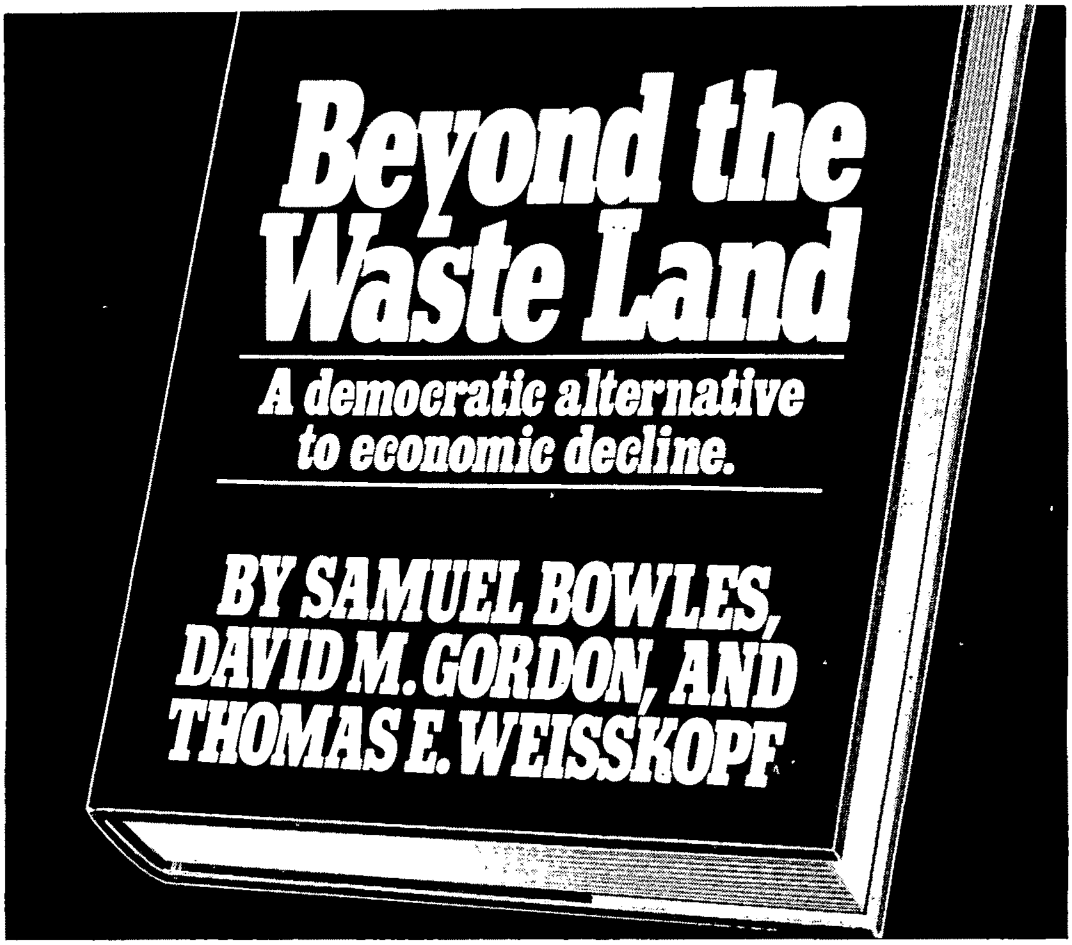
"**His chapter on inflation is a tour de force** of analysis and lucidity...The rigor and breadth of his analysis is testimony to the richness that still remains in economic theory."
–JEFFREY MADRICK, *New York Times**



$16.95, now at
your bookstore

RANDOM
HOUSE

Photo: Susan Koenig


Beyond the Waste Land
A democratic alternative to economic decline.
BY SAMUEL BOWLES, DAVID M. GORDON, AND THOMAS E. WEISSKOPF

The alternative Reagan asked for– "by the three most interesting economists of the left."–*John Kenneth Galbraith*

"By and large, the economics profession has not developed any powerfully convincing diagnosis of the stagflation which has afflicted the United States and world economies the past decade. Treatments based on reigning orthodox theories appear to be worse than the disease. So both economists and other concerned citizens should take seriously the refreshing and original explanations of our problems offered by Bowles, Gordon, and Weisskopf, and their detailed proposals for solution. It is good to see left-wing political economy come alive."—JAMES TOBIN, Sterling Professor of Economics, Yale University, Nobel Laureate in Economics.

"A novel explanation for the turndown in productivity based on worker reactions and corporate control. It is well worth study and consideration."—KENNETH J. ARROW, Joan Kenney Professor of Economics, Stanford University, Nobel Laureate in Economics.

ANCHOR PRESS/DOUBLEDAY

just published • just published • just published • just published • just published

**Editorial Board:**

**Prof. H. Akuto,**
University of Tokyo, Japan
**Prof. K. Arrow,**
Stanford University, U.S.A.
**Dr. E. E. Bailey,**
Civil Aeronautics Board, U.S.A.
**Dr. M. Benedetti,**
S.T.E.T., Italy
**Prof. Y. M. Braunstein,**
Brandeis University, U.S.A.
**Mr. R.J. Chapuis,**
C.C.I.T.T., Switzerland
**Prof. A. J. Flowerdew,**
University of Kent, U.K.
**Dr. H. P. Gassmann,**
O.E.C.D., France
**Dr. C.L. Hilewick,**
A.T. & T., U.S.A.
**Prof. H. Inose,***
University of Tokyo, Japan
**Prof. M.D. Intriligator,**
University of California (L.A.), U.S.A.
**Dr. C. Jacobaeus,**
Ericsson, Sweden
**Prof. C. Jonscher,**
M.I.T., U.S.A.
**Prof. M. Jussawalla,**
Communication Institute, U.S.A.
**Prof. P. Laconte,**
University of Louvain, Belgium
**Prof. D. M. Lamberton,***
University of Queensland, Australia
**Prof. P. Mac Avoy,**
Yale University, U.S.A.
**Prof. J.J. McCall,***
University of California (L.A.), U.S.A.
**Dr. B. Mitchell,**
Rand Corporation, U.S.A.
**Dr. J. Müller,**
Deutsches Inst. für Wirtschaftsforschung, F.R.G.
**Dr. M. Nagai,**
United Nations University, Japan
**Prof. R.G. Noll,***
California Inst. of Technology, U.S.A.
**Dr. S. Okita,**
Ministry of Foreign Affairs, Japan
**Prof. H. Oniki,**
Osaka University, Japan
**Dr. J. C. Panzar,**
Bell Laboratories, U.S.A.
**Prof. M. Spence,***
Harvard University, U.S.A.
**Prof. N. Takasaki,**
Res. Inst. of Telecommunications and Economics, Japan
**Prof. R. M. Townsend,**
Carnegie-Mellon University, U.S.A.
**Dr. M. Tyler,**
Communication Studies and Planning, U.S.A.
**Prof. K. Valaskakis,**
GAMMA, Canada
**Dr. J. Voge,***
Idate, France
**Prof. C. C von Weizsäcker,***
University of Bern, Switzerland
**Mr. K. Yukihiro,**
Res. Inst. of Telecommunications and Economics, Japan

*Supervisory and Policy Board

**North-Holland**

# INFORMATION ECONOMICS AND POLICY

**A NEW international and inter-disciplinary journal for:**

- **RESEARCHERS**
- **EXPERT CONSULTANTS**
- **POLICY MAKERS**

**in the fields of:**

- **INFORMATION ECONOMICS**
- **(TELE) COMMUNICATIONS**
- **MEDIA**

INFORMATION ECONOMICS AND POLICY provides an international forum for research papers, short contributions and surveys with (tele) communications economics and policy as its core, including related issues on information economics and media policy.

The contributions should be of a high quality standard and be accessible in language and presentation to an international, interdisciplinary audience. The following are typical examples of what will be included:

- Policy issues related to economics of the media and information;
- Economic and policy issues in the area of information goods and services;
- The economic of information in public and private decision making and strategy;
- Economic analysis of primary and secondary information systems;
- Contribution of (tele) communications and office automation to "white collar" productivity;
- Economics and policy of transfer of data and knowhow and transborder information systems;
- International studies of policy and economics of information systems;
- Economics of information networks and the analysis of new information technology, e.g. direct satellite broadcasting, teleconferencing, viewdata;
- Economics of information in markets, trade and competition;
- Economics of advertising and promotion.

## Publication Information:

INFORMATION ECONOMICS AND POLICY is scheduled for quarterly publication comprising ± 90 pages per issue. Volume 1, number 1, started publication in July 1983.
Subscription price for volume 1 in 4 issues: Dfl. 200.00/US $80.00, including postage.

.Please mail this coupon (or photocopy) to:

**Elsevier Science Publishers**
Subscription Order Department
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

For customers in the USA/Canada:
**Elsevier Science Publishing Co., Inc.**
Journal Information Center
P.O. Box 1663, Grand Central Station,
New York, NY 10163, U.S.A.

*Please send me a free specimen copy of:*

**INFORMATION ECONOMICS AND POLICY**

Name ______________________________

Professional Address ______________________________

______________________________

______________________________

NH/MAN/JRL/6095a

# THE ECONOMICS OF SUPPLY AND DEMAND

***Lawrence Klein***

*Winner of the 1980 Nobel Prize in Economics*

*The Economics of Supply and Demand* takes a critical look at the foundation of Reaganomics—supply-side theory—and at the Keynesian doctrines it replaced.

Arguing that neither supply analysis alone nor demand analysis alone explains what is happening in the economy, Klein illustrates the interaction of supply- and demand-side forces and the importance of that interaction in devising effective economic policies.

Challenging concepts such as monetarism and rational expectations, the author advocates an unconventional, almost macroeconomic interpretation of supply-side theory and recommends major shifts in the direction of economic policy. $17.50

**The Johns Hopkins University Press, Baltimore, Maryland 21218**

*New!*

# SUPREME COURT ECONOMIC REVIEW

Editor: PETER H. ARANSON, *Law and Economics Center, Emory University*

The power of the United States Supreme Court to affect economic, political, and social life requires no emphasis. But until now Court-watchers have had little coherent theory in seeking to understand the basis for, and the consequences of, the Court's decisions and opinions. Now, leading scholars use the new discipline of Law and Economics to analyze key cases from each Term of the Court in *The Supreme Court Economic Review*.

An annual publication of the Emory University Law and Economics Center, the most distinguished institution of its kind, the *Review* focuses interdisciplinary scholarship on opinions of the Court that will significantly affect the arrangements of American society—especially economic organization, property, and contractual rights and obligations. *Volume 1* covers the 1980 Term of the Court.

*Order from*

**MACMILLAN Publishing Company**

200D BROWN STREET, RIVERSIDE, NJ 08075

**Contents: Volume 1**

The Regulation Machine, James E. Krier, University of California, Los Angeles.

Democracy and Representation: A Reconciliation of *Ball v. James* and *Reynolds v. Sims*, William H. Riker, University of Rochester.

Incidence Analysis and the Supreme Court: An Examination of Four Cases from the 1980 Term, Charles E. McLure, Jr., Hoover Institution, Stanford University.

Prohibition of Billboard Advertising: An Economic Analysis of the *Metromedia* Decision, Keith B. Leffler, University of Washington.

Efficiency and Vicarious Liability for Punitive Damages: Economic Implications of *City of Newport v. Facts Concerts, Inc.*, Dorsey D. Ellis, Jr., University of Iowa.

Regulating the Content and Volume of Litigation: An Economic Analysis, George L. Priest, Yale University.

*County of Washington v. Gunther*: Legal and Economic Considerations for Resolving Sex-Based Wage Discrimination Cases, Cotton M. Lindsay and Charles A. Shanor, both, Emory University.

Decision Sharing and Expropriable Specific Quasi-Rents: A Theory of *First National Maintenance Corporation v. NLRB*, Armen A. Alchian, University of California, Los Angeles.

*320 pages • Cloth • ISBN 0-02-918160-7 • $29.95*

**DUKE PRESS POLICY STUDIES** **DPPS**

**What Role for Government?**
**Lessons from Policy Research**
*Edited by* Richard J. Zeckhauser, *Harvard, and* Derek Leebaert, *Harvard* $37.75

**The Making of Federal Coal Policy**
Robert H. Nelson, *Office of Policy Analysis, U.S. Dept. of Interior* September, ca.$35.00

---

**Published in cooperation with the Council of State Planning Agencies**
*Edited by* Michael Barker, *Gallatin Institute*

**State Taxation Policy and Economic Growth** $31.75

**Financing State and Local Economic Development** $37.50

**Rebuilding America's Infrastructure**
**An Agenda for the 1980s** September, ca.$30.00

**State Employment Policy in Hard Times** September, ca.$25.00

**America in Ruins**
**The Decaying Infrastructure**
Pat Choate, *TRW, Inc., and* Susan Walter, *General Electric* paper, $9.75

---

**The World Tin Market**
**Political Pricing and Economic Competition**
William L. Baldwin, *Dartmouth* September, ca.$45.00

**Modeling Growing Economies in Equilibrium and Disequilibrium**
*Edited by* Allen C. Kelley, *Duke;* Warren C. Sanderson, *SUNY—Stony Brook; and* Jeffrey G. Williamson, *Wisconsin* $37.50

---

NEW EDITION OF CLASSIC TEXT

**The Growth of Economic Thought**
**Revised and Expanded Edition**
Henry William Spiegel, *Catholic University* cloth, $37.50; paper, $22.50

---

**Duke University Press**
**6697 College Station / Durham, North Carolina 27708**

# HARCOURT BRACE JOVANOVICH

Publishers of Baumol and Blinder's
**PRINCIPLES OF ECONOMICS, 2nd Ed.,**
also provide these other outstanding economics texts.

### INTERNATIONAL ECONOMICS: A POLICY APPROACH, 4th Ed.

**Mordechai Kreinin,**
*Michigan State University*

This comprehensive introduction to international finance and trade places a special emphasis on international economic policy and policy analysis. Nine appendixes offer technical and theoretical discussions.
432 pages, 1983.

### HISTORY OF THE AMERICAN ECONOMY, 5th Ed.

**Gary M. Walton,**
*University of California, Davis*
**Ross M. Robertson,**
*Late of Indiana University*

Maintaining a careful balance between descriptive and analytical economic history, Walton provides a comprehensive and highly readable account of the development of American commerce.
692 pages, 1983.

### THE ECONOMIC TRANSFORMATION OF AMERICA, 1600 TO THE PRESENT, 2nd Ed.

**Robert L. Heilbroner,**
*New School for Social Research*
**Aaron Singer,**
*Manhattan College*

This extensively revised Second Edition discusses the process of economic development in America from its founding to the present. Material on the Colonial economic struggle has been added, and the examination of post-war American economy has been updated.
368 pages (tentative), 1984.

### MONEY AND THE ECONOMY, 5th Ed.

**John J. Klein,**
*Georgia State University*

Providing a balanced analysis and coverage of the classical, Keynesian, monetarist, and neo-Keynesian positions, Klein explains financial institutions and their role in the economy, and monetary theory and policy.
538 pages, 1982.

### CONTESTABLE MARKETS AND THE THEORY OF INDUSTRY STRUCTURE

**William J. Baumol,**
*Princeton University and New York University*
**John C. Panzar,**
*Bell Laboratories*
**Robert D. Willig,**
*Princeton University*

Using sophisticated graphical and mathematical analysis, the authors present an integrative theory of the determinants of prices and industrial structure in contestable markets.
510 pages, 1982.
No examination copies; purchase only.

### FINANCIAL INSTITUTIONS, FINANCIAL MARKETS, AND MONEY

**Herbert M. Kaufman,**
*Arizona State University*

Kaufman describes and analyzes financial markets and institutions in a changing environment, providing careful discussions of the role of regulation and innovation in the financial system and the interrelationship between institutions and the marketplaces in which they participate.
560 pages, 1983.

HBJ

**Harcourt Brace Jovanovich Publishers,** 1250 Sixth Avenue, San Diego, CA 92101

# Forecasting and Statistics

**ELF**

- ☐ factor analysis
- ☐ discriminant analysis
- ☐ stepwise regression with residuals and forecasting
- ☐ crosstabs
- ☐ uses Basic commands for transformations
- ☐ hi-res scattergram
- ☐ means, standard deviations, etc.
- ☐ correlation coefficients
- ☐ 1 and 2 way ANOVA
- ☐ report writer
- ☐ t-test on means
- ☐ database manager

For Apple and CP/M

**TWG/ARIMA**

- ☐ Box-Jenkins package
- ☐ seasonal and non-seasonal models
- ☐ Box-Cox transformations
- ☐ identification, estimation and forecasting
- ☐ statistics, SSE grid, confidence intervals, etc.
- ☐ database manager

**Now available for CP/M!**

**Send me information on:** ☐ ELF ☐ TWG/ARIMA.
I use: ☐ Apple ☐ CP/M

Name: ____________________

Organization/Affiliation: ____________________

Address: ____________________

City: __________ State: __________ Zip: __________

Mail to: **THE WINCHENDON GROUP,** 3907 Lakota Rd., PO Box 10114,
Alexandria, VA 22310, or phone 1-800-431-1953 x831; in New York
1-800-942-1935 x831

## THE WINCHENDON GROUP

# Oxford

## The Economics of Inequality
**Second Edition**

A.B. ATKINSON, *London School of Economics and Political Science.* This second edition of a very successful book on the distribution of income and wealth has been substantially rewritten and updated. *On the first edition:* "Throughout the argument is trenchant, technically expert and yet crystal clear. One really does get the feeling that one has here an author who is thoroughly at home in the field in which he is writing."—*Times Literary Supplement*

1983 296 pp.; 29 illus.
cloth $27.50 paper $11.50

## Pay Policies for the Future

Edited by DEREK ROBINSON and KEN MAYHEW, both of the *University of Oxford.* A wide range of specialists in the areas of arbitration and industrial relations contribute their ideas concerning the options open to governments in their attempts to influence pay.

1983 144 pp.; 2 tables, 1 fig. paper $9.95

## The Economic Analysis of Technological Change

PAUL STONEMAN, *University of Warwick.* This book presents a comprehensive view of the economic approach to the analysis of technical change. The author examines the generation of new technology, surveys the literature on technological diffusion, and investigates the impacts of new technology on today's economy.

October 1983 256 pp.; 22 illus.
cloth $37.50 paper $14.95

## The Foundations of Macroeconomic and Monetary Theory

P.J.N. SINCLAIR, *University of Oxford.* This analysis of Keynesian economics examines the key elements that form the basis of the real and monetary sides of the macroeconomy. The book also surveys models of the aggregate economy and focuses upon the longer-term issues of monetary policy and growth.

October 1983 192 pp.; 6 tables, 83 figs.
cloth $42.00 paper $18.95

## Marxism and the Metropolis
**New Perspectives in Urban Political Economy**
**Second Edition**

Edited by WILLIAM K. TABB, *Queens College, City University of New York,* and LARRY SAWERS, *American University.* This volume brings together papers in the field of urban political economy in which the contributors analyze the spatial structure of the city, suburbanization, financing urban governments, urban redevelopment and transportation, and urban housing.

November 1983 384 pp.; 4 maps
cloth $24.95 paper $11.95

## Inequalising Trade?
**A Study of Traditional North/South Specialisation in the Context of Terms of Trade Concepts**

JOHN SPRAOS, *University College, London.* A challenging contribution to a crucial and controversial subject, this book treats an issue that has dominated economic relations across the rich/poor, North/South divide for more than twenty-five years.

1983 192 pp.; 9 illus., 7 tables $32.50

## Financial Ratios for Monitoring Corporate Sickness

L.C. GUPTA, *Xavier Labour Relations Institute, Jamshedpur, India.* The core of this study deals with the task of empirically testing a wide variety of profitability and balance sheet ratios to identify which can offer the most meaningful warning signals and be built into a predictive system.

1983 160 pp.; 4 charts $16.95

## The Origins of Trade Union Power

HENRY PHELPS-BROWN, *London School of Economics.* This book throws light on the sources and characteristics of trade unionism by singling out some of the critical moments and vital forces in its history. It examines particularly the relations between trade unions and the law and the endeavors of governments to exert controls.

September 1983 234 pp.; 1 graph $32.50

# World Development Report 1983

## THE WORLD BANK

The *World Development Report* is recognized worldwide as essential reading for economists, students, planners and policy makers, bankers and businessmen, journalists, and development professionals of all kinds. *World Development Report 1983* is the sixth in an annual series of reports from The World Bank. It reviews how the developing countries have fared in the face of international economic trends and emphasizes the potential for increasing development performance through better resource management. The key issue raised is that of the limitations on governments' capacities to intervene and administer. Finance development and debt management are discussed as well as the management of public bureaucracies and the impact of foreign aid. The *Report* features full-color charts, an annotated bibliography, and an annex of world development indicators citing the most up-to-date information on developing nations.

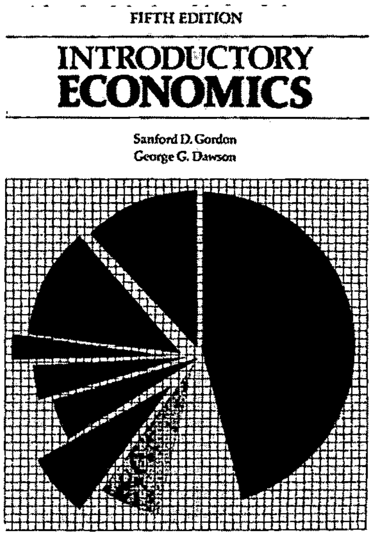
CONTENTS:
Overview
The Prolonged Recession
The Outlook for Developing Countries
The Search for Efficiency
The Role of the State
Pricing for Efficiency
National Economic Management
Managing State-owned Enterprises
Project and Program Management
Managing the Public Service
Reorienting Government
Concluding Themes

Now Available 216 pp.; charts, graphs
cloth $20.00 paper $8.00

**OXFORD UNIVERSITY PRESS**
**200 Madison Avenue, New York, New York 10016**

*Announcing A New Edition*
*For Your One-Term Course . . .*

*Now available in paperback and hardback!*



FIFTH EDITION

INTRODUCTORY
ECONOMICS

Sanford D. Gordon
George G. Dawson

**INTRODUCTORY ECONOMICS**
**Fifth Edition**

**Sanford D. Gordon,** New York State Council on Economic Education, Russell Sage College
**George G. Dawson,** Empire State College, State University of New York
1984 Hardback and Paperback 427 pages
Study Guide/Instructor's Guide

Practical, real-world problems are the focus of this lively and accessible text for the one-term course. Based on an analytical-problem-solving approach, the book's proven teaching and learning system features four broad review models — demand and supply, long-run equilibrium, pricing factors of production, and natural income determination. Within this conceptual framework, students are introduced to microeconomic and macroeconomic theory with applications to current problems and issues, such as the economics of pollution, small business, wage and price controls, and inflation.

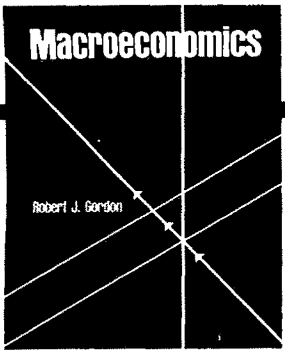
Graphic and mathematical analyses are simply and clearly presented. Highly successful learning aids include: chapter objectives, summaries, highlights, key terms and names, review questions and problems, a glossary of terms, and a new appendix on economics in careers.

**Highlights**
- A new step-by-step orientation into the role and meaning of tables and graphs in economics
- An update on the most recent developments in labor relations, such as the phenomenon of "give-backs" and the growing interest in quality circles
- An expanded analysis on the impact of OPEC
- New information on the dramatic changes occurring in our system of money and banking
- Completely revised chapter on the effects of the recession and the comparison between supply-side economics and the demand-side approach of Keynes
- Updated chapter on environmental problems and energy
- New information on our relations with the Common Market, the dollar's role in the world economy, and the third world's difficulty in paying its external debts

*For details or sample copies, call us toll free: 800-225-1388*
*In Massachusetts, call collect: 617-862-6650, ext. 1344*

***From D. C. Heath!***

**D. C. HEATH AND COMPANY**
HEATH
College Division
125 Spring Street
Lexington, MA 02173



New edition of a proven leader in Intermediate Macro . . .

**MACROECONOMICS**
*Third Edition*
**Robert J. Gordon**
Northwestern University

**"The revisions are nothing short of superb."**
William R. Hart, *Miami University*

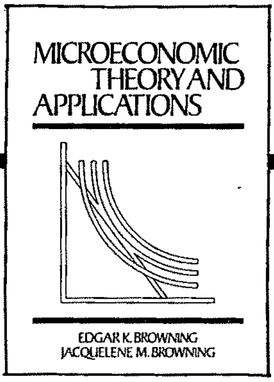
**"The development of the labor market in the manuscript is far superior to the second edition in clarity and coherence."**
Bruce Bolnick, *University of North Carolina*

**"In my judgement, it is the best blend of theory, policy and evidence at the intermediate level."**
Jerome L. McElroy, *St. Mary's College*

**"The thing I like best about Gordon's text is his integration of theory with explanations of real economic events."**
Alan Rabin, *University of Tennessee*

**"I don't think I've read a better summary of the key propositions that divide the monetarists and Keynesians."**
John Flanders, *Central Methodist University*

In thoroughly updating his bestselling text, Robert J. Gordon paid particular attention to the chapters on inflation and unemployment (Part III) — the result according to prepublication reviewers is "right on the mark." Of course, MACROECONOMICS continues to focus on key policy issues, to provide up-to-the-minute theoretical coverage, and to examine real world economic events through intriguing case studies. The breadth and quality of the supplementary materials — including Student Workbook, Instructor's Manual, and "The Gordon Update" (published twice yearly) — make this the state of the art teaching package.
*cloth/c. 720 pages*
***Available for examination in November!***



A growing new success in Micro — adopted by 200 schools across the country in less than a year

**MICROECONOMIC THEORY AND APPLICATIONS**
**Edgar K. Browning**
University of Virginia
**Jacquelene M. Browning**
Sweet Briar College

**"A superb blend of issues, analytics, and evidence."**
David E. R. Gay, *University of Arkansas*

**". . . there is a clarity of line or thought in the presentation of theory which is missing in many existing texts."**
Timothy J. Gronberg, *Texas A & M University*

**"I can't remember any book that has taught me more interesting things in economics than this one. I am overwhelmed by the depth and originality of their presentation."**
Roy Van Til, *Bentley College*

**"The treatment is masterful in combining fundamental principles with practical and interesting applications."**
William McEachern, *University of Connecticut*

**"The concepts are presented in perhaps the most lucid manner I have found in any text."**
Edward R. Kittrell, *Northern Illinois University*

Fulfilling the expectations of its enthusiastic reviewers, MICROECONOMIC THEORY AND APPLICATIONS has already become one of the leading texts in the field. It combines solid theory, dynamic applications, and illuminating, two-color graphics with incomparably clear and complete coverage of core topics. Instructive and engaging applications are integrated throughout each theory chapter, and four separate chapters (on consumer choice theory, the competitive model, monopoly theory, and input market analysis) are devoted entirely to applications.
*cloth/604 pages/1983*

**LITTLE, BROWN AND COMPANY** 34 Beacon Street, Boston, Massachusetts 02106

# *New from The World Bank*

## Rural Financial Markets in Developing Countries

THEIR USE AND ABUSE

***edited by J. D. von Pischke, Dale W. Adams, and Gordon Donald***

A wide-ranging exploration of the problems plaguing rural financial markets in developing countries. The authors discuss the developmental role of finance, credit and finance in farm households, the operation of both urban-based and local financial institutions in rural areas, and national policies affecting rural financial markets.

The authors view credit as part of financial markets, rather than as an adjunct to agriculture or rural development — a perspective that challenges widely held notions about the benefits of low interest rates to rural development, the ability of small farmers to save, and the fairness of informal types of moneylending. **$32.50**

## Food Policy Analysis

***C. Peter Timmer, Walter P. Falcon, and Scott R. Pearson***

*Food Policy Analysis* describes how food systems — the mechanisms by which farm commodities are produced, transformed into food, and sold to consumers — function in different environments and discusses policy interventions that can give the poor greater access to food.

The authors' analysis has three components: the behavior of food consumers and producer households; the effects of macroeconomic forces on the performance of the food system; and the role of markets, both domestic and international, in linking household issues in the micro sector to policy issues in the macro economy. **$25.00** hardcover, **$12.95** paperback

## International Finance for Food Security

***Barbara Huddleston, D. Gale Johnson,***
***Shlomo Reutlinger, and Alberto Valdés***

An examination of the economic justification for creating an international facility to finance a portion of the food imports to developing countries when special needs arise. Using a stochastic simulation model, the authors analyze the likely destabilizing effect and benefits of the financial facility for different policy scenarios and countries. They also evaluate the possible contribution of such a facility and assess the merits and probable effects of a facility that has been adopted by the International Monetary Fund. **$15.00**

***Published for The World Bank by***
**The Johns Hopkins University Press**
Baltimore, Maryland 21218

Begin making plans to attend the

# *Ninety-Sixth*

# Annual Meeting of The American Economic Association

**(in Conjunction with Allied Social Science Associations)**

to be held in

# SAN FRANCISCO, CA.

**Dec. 28-30, 1983**

---

The Employment Center opens Tuesday, December 27.

---

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

---

The 1984 meeting will be held in Dallas, TX, December 28-30.

---

# The '84 outlook . . . it never looked so good!

third edition

## MACROECONOMICS

**Paul Wonnacott**, University of Maryland

- Workbook/Study Guide
- Instructor's Manual

sixth edition

## MANAGERIAL ECONOMICS

Text, Problems, and Short Cases

**K. K. Seo**, University of Hawaii-Manoa

- Study Guide
- Instructor's Manual

fourth edition

## THE NEW WORLD OF ECONOMICS

Explorations into the Human Experience

**Gordon Tullock**, Virginia Polytechnic and State University

**Richard B. McKenzie**, Clemson University

- paperback

fifth edition

## LABOR ECONOMICS

Wages, Employment, Trade Unionism, and Public Policy

**F. Ray Marshall**, University of Texas at Austin

**Vernon M. Griggs, Jr.**, Cornell Univeristy

**Allan G. King**, University of Texas at Austin

*When requesting examination copies of these outstandingtexts for adoption consideration, please indicate course title and text presently used.*

**Richard D.**
**IRWIN** • Inc.
Homewood, Illinois 60430

50 Years of Excellence 1933-1983

29 OCT 1984